3
Network Design

Before purchasing equipment or deciding on a hardware platform, you should have a clear idea of the nature of your communications problem. Most likely, you are reading this book because you need to connect computer networks together in order to share resources and ultimately reach the larger global Internet. The network design you choose to implement should fit the communications problem you are trying to solve. Do you need to connect a remote site to an Internet connection in the center of your campus? Will your network likely grow to include several remote sites? Will most of your network components be installed in fixed locations, or will your network expand to include hundreds of roaming laptops and other devices?

In this chapter, we will begin with a review of the networking concepts that define TCP/IP, the primary family of networking protocols currently used on the Internet. We will then see examples of how other people have built wireless networks to solve their communication problems, including diagrams of the essential network structure. Finally, we will present several common methods for getting your information to flow efficiently through your network and on to the rest of the world.

Networking 101

TCP/IP refers to the suite of protocols that allow conversations to happen on the global Internet. By understanding TCP/IP, you can build networks that will scale to virtually any size, and will ultimately become part of the global Internet.

If you are already comfortable with the essentials of TCP/IP networking (including addressing, routing, switches, firewalls, and routers), you may want

to skip ahead to **Designing the Physical Network** on **Page 51**. We will now review the basics of Internet networking.

Introduction

Venice, Italy is a fantastic city to get lost in. The roads are mere foot paths that cross water in hundreds of places, and never go in a simple straight line. Postal carriers in Venice are some of the most highly trained in the world, specializing in delivery to only one or two of the six *sestieri* (districts) of Venice. This is necessary due to the intricate layout of that ancient city. Many people find that knowing the location of the water and the sun is far more useful than trying to find a street name on a map.

Figure 3.1: Another kind of network mask.

Imagine a tourist who happens to find papier-mâché mask as a souvenir, and wants to have it shipped from the studio in S. Polo, Venezia to an office in Seattle, USA. This may sound like an ordinary (or even trivial) task, but let's look at what actually happens.

The artist first packs the mask into a shipping box and addresses it to the office in Seattle, USA. They then hand this off to a postal employee, who attaches some official forms and sends it to a central package processing hub for international destinations. After several days, the package clears Italian customs and finds its way onto a transatlantic flight, arriving at a central import processing location in the U.S. Once it clears through U.S. customs, the package is sent to the regional distribution point for the northwest U.S., then on to the Seattle postal processing center. The package eventually makes its way onto a delivery van which has a route that brings it to the proper address, on the proper street, in the proper neighborhood. A clerk at the office

accepts the package and puts it in the proper incoming mail box. Once it arrives, the package is retrieved and the mask itself is finally received.

The clerk at the office in Seattle neither knows nor cares about how to get to the *sestiere* of S. Polo, Venezia. His job is simply to accept packages as they arrive, and deliver them to the proper person. Similarly, the postal carrier in Venice has no need to worry about how to get to the correct neighborhood in Seattle. His job is to pick up packages from his local neighborhood and forward them to the next closest hub in the delivery chain.

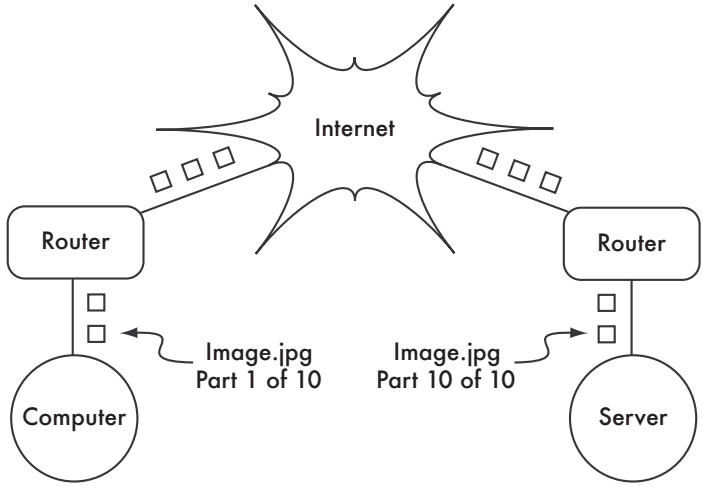


Figure 3.2: Internet networking. Packets are forwarded between routers until they reach their ultimate destination.

This is very similar to how Internet routing works. A message is split up into many individual ***packets***, and are labeled with their source and destination. The computer then sends these packets to a ***router***, which decides where to send them next. The router needs only to keep track of a handful of routes (for example, how to get to the local network, the best route to a few other local networks, and one route to a gateway to the rest of the Internet). This list of possible routes is called the ***routing table***. As packets arrive at the router, the destination address is examined and compared against its internal routing table. If the router has no explicit route to the destination in question, it sends the packet to the closest match it can find, which is often its own Internet gateway (via the ***default route***). And the next router does the same, and so forth, until the packet eventually arrives at its destination.

Packages can only make their way through the international postal system because we have established a standardized addressing scheme for packages. For example, the destination address must be written legibly on the front of the package, and include all critical information (such as the recipient's name,

street address, city, country, and postal code). Without this information, packages are either returned to the sender or are lost in the system.

Packets can only flow through the global Internet because we have agreed on a common addressing scheme and protocol for forwarding packets. These standard communication protocols make it possible to exchange information on a global scale.

Cooperative communications

Communication is only possible when the participants speak a common language. But once the communication becomes more complex than a simple conversation between two people, protocol becomes just as important as language. All of the people in an auditorium may speak English, but without a set of rules in place to establish who has the right to use the microphone, the communication of an individual's ideas to the entire room is nearly impossible. Now imagine an auditorium as big as the world, full of all of the computers that exist. Without a common set of communication protocols to regulate when and how each computer can speak, the Internet would be a chaotic mess where every machine tries to speak at once.

People have developed a number of communications frameworks to address this problem. The most well-known of these is the ***OSI model***.

The OSI model

The international standard for Open Systems Interconnection (OSI) is defined by the document ISO/IEC 7498-1, as outlined by the International Standards Organization and the International Electrotechnical Commission. The full standard is available as publication "ISO/IEC 7498-1:1994," available from *http://standards.iso.org/ittf/PubliclyAvailableStandards/*.

The OSI model divides network traffic into a number of ***layers***. Each layer is independent of the layers around it, and each builds on the services provided by the layer below while providing new services to the layer above. The abstraction between layers makes it easy to design elaborate and highly reliable ***protocol stacks***, such as the ubiquitous ***TCP/IP*** stack. A protocol stack is an actual implementation of a layered communications framework. The OSI model doesn't define the protocols to be used in a particular network, but simply delegates each communications "job" to a single layer within a well-defined hierarchy.

While the ISO/IEC 7498-1 specification details how layers should interact with each other, it leaves the actual implementation details up to the manufacturer. Each layer can be implemented in hardware (more common for lower layers) or software. As long as the interface between layers adheres to

the standard, implementers are free to use whatever means are available to build their protocol stack. This means that any given layer from manufacturer A can operate with the same layer from manufacturer B (assuming the relevant specifications are implemented and interpreted correctly).

Here is a brief outline of the seven-layer OSI networking model:

Layer	Name	Description
7	Application	The ***Application Layer*** is the layer that most network users are exposed to, and is the level at which human communication happens. HTTP, FTP, and SMTP are all application layer protocols. The human sits above this layer, interacting with the application.
6	Presentation	The ***Presentation Layer*** deals with data representation, before it reaches the application. This would include MIME encoding, data compression, formatting checks, byte ordering, etc.
5	Session	The ***Session Layer*** manages the logical communications session between applications. NetBIOS and RPC are two examples of a layer five protocol.
4	Transport	The ***Transport Layer*** provides a method of reaching a particular service on a given network node. Examples of protocols that operate at this layer are TCP and UDP. Some protocols at the transport layer (such as TCP) ensure that all of the data has arrived at the destination, and is reassembled and delivered to the next layer in the proper order. UDP is a "connectionless" protocol commonly used for video and audio streaming.
3	Network	IP (the Internet Protocol) is the most common ***Network Layer*** protocol. This is the layer where routing occurs. Packets can leave the link local network and be retransmitted on other networks. Routers perform this function on a network by having at least two network interfaces, one on each of the networks to be interconnected. Nodes on the Internet are reached by their globally unique IP address. Another critical Network Layer protocol is ICMP, which is a special protocol which provides various management messages needed for correct operation of IP. This layer is also sometimes referred to as the ***Internet Layer***.

Layer	Name	Description
2	Data Link	Whenever two or more nodes share the same physical medium (for example, several computers plugged into a hub, or a room full of wireless devices all using the same radio channel) they use the ***Data Link Layer*** to communicate. Common examples of data link protocols are Ethernet, Token Ring, ATM, and the wireless networking protocols (802.11a/b/g). Communication on this layer is said to be link-local, since all nodes connected at this layer communicate with each other directly. This layer is sometimes known as the ***Media Access Control*** (***MAC***) layer. On networks modeled after Ethernet, nodes are referred to by their ***MAC address.*** This is a unique 48 bit number assigned to every networking device when it is manufactured.
1	Physical	The ***Physical Layer*** is the lowest layer in the OSI model, and refers to the actual physical medium over which communications take place. This can be a copper CAT5 cable, a fiber optic bundle, radio waves, or just about any other medium capable of transmitting signals. Cut wires, broken fiber, and RF interference are all physical layer problems.

The layers in this model are numbered one through seven, with seven at the top. This is meant to reinforce the idea that each layer builds upon, and depends upon, the layers below. Imagine the OSI model as a building, with the foundation at layer one, the next layers as successive floors, and the roof at layer seven. If you remove any single layer, the building will not stand. Similarly, if the fourth floor is on fire, then nobody can pass through it in either direction.

The first three layers (Physical, Data Link, and Network) all happen "on the network." That is, activity at these layers is determined by the configuration of cables, switches, routers, and similar devices. A network switch can only distribute packets by using MAC addresses, so it need only implement layers one and two. A simple router can route packets using only their IP addresses, so it need implement only layers one through three. A web server or a laptop computer runs applications, so it must implement all seven layers. Some advanced routers may implement layer four and above, to allow them to make decisions based on the higher-level information content in a packet, such as the name of a website, or the attachments of an email.

The OSI model is internationally recognized, and is widely regarded as the complete and definitive network model. It provides a framework for manufac-

turers and network protocol implementers that can be used to build networking devices that interoperate in just about any part of the world.

From the perspective of a network engineer or troubleshooter, the OSI model can seem needlessly complex. In particular, people who build and troubleshoot TCP/IP networks rarely need to deal with problems at the Session or Presentation layers. For the majority of Internet network implementations, the OSI model can be simplified into a smaller collection of five layers.

The TCP/IP model

Unlike the OSI model, the TCP/IP model is not an international standard and its definitions vary. Nevertheless, it is often used as a pragmatic model for understanding and troubleshooting Internet networks. The vast majority of the Internet uses TCP/IP, and so we can make some assumptions about networks that make them easier to understand. The TCP/IP model of networking describes the following five layers:

Layer	Name
5	Application
4	Transport
3	Internet
2	Data Link
1	Physical

In terms of the OSI model, layers five through seven are rolled into the topmost layer (the Application layer). The first four layers in both models are identical. Many network engineers think of everything above layer four as "just data" that varies from application to application. Since the first three layers are interoperable between virtually all manufacturers' equipment, and layer four works between all hosts using TCP/IP, and everything above layer four tends to apply to specific applications, this simplified model works well when building and troubleshooting TCP/IP networks. We will use the TCP/IP model when discussing networks in this book.

The TCP/IP model can be compared to a person delivering a letter to a downtown office building. The person first needs to interact with the road itself (the Physical layer), pay attention to other traffic on the road (the Data Link layer), turn at the proper place to connect to other roads and arrive at the correct address (the Internet layer), go to the proper floor and room num-

ber (the Transport layer), and finally give it to a receptionist who can take the letter from there (the Application layer). Once they have delivered the message to the receptionist, the delivery person is free to go on their way.

The five layers can be easily remembered by using the mnemonic "**P**lease **D**on't **L**ook **I**n **T**he **A**ttic," which of course stands for "**P**hysical / **D**ata **L**ink / **I**nternet / **T**ransport / **A**pplication."

The Internet protocols

TCP/IP is the protocol stack most commonly used on the global Internet. The acronym stands for ***Transmission Control Protocol*** (***TCP***) and ***Internet Protocol*** (***IP***), but actually refers to a whole family of related communications protocols. TCP/IP is also called the ***Internet protocol suite***, and it operates at layers three and four of the TCP/IP model.

In this discussion, we will focus on version four of the IP protocol (IPv4) as this is now the most widely deployed protocol on the Internet.

IP Addressing

In an IPv4 network, the address is a 32-bit number, normally written as four 8-bit numbers expressed in decimal form and separated by periods. Examples of IP addresses are 10.0.17.1, 192.168.1.1, or 172.16.5.23.

If you enumerated every possible IP address, they would range from 0.0.0.0 to 255.255.255.255. This yields a total of more than four billion possible IP addresses (255 x 255 x 255 x 255 = 4,228,250,625); although many of these are reserved for special purposes and should not be assigned to hosts. Each of the usable IP addresses is a unique identifier that distinguishes one network node from another.

Interconnected networks must agree on an IP addressing plan. IP addresses must be unique and generally cannot be used in different places on the Internet at the same time; otherwise, routers would not know how best to route packets to them.

IP addresses are allocated by a central numbering authority that provides a consistent and coherent numbering method. This ensures that duplicate addresses are not used by different networks. The authority assigns large blocks of consecutive addresses to smaller authorities, who in turn assign smaller consecutive blocks within these ranges to other authorities, or to their customers. These groups of addresses are called sub-networks, or ***subnets*** for short. Large subnets can be further subdivided into smaller subnets. A group of related addresses is referred to as an ***address space***.

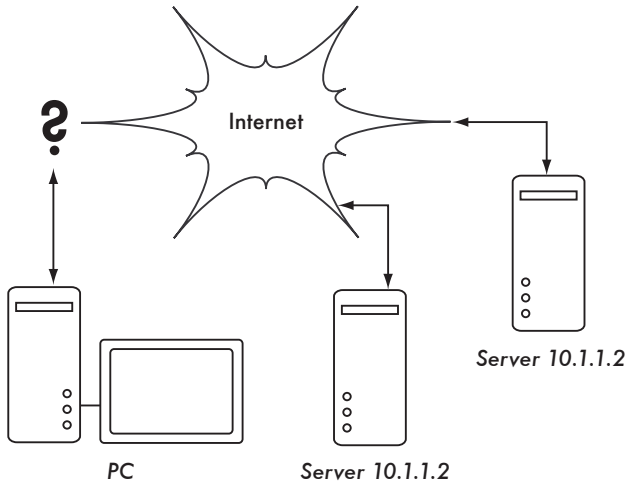


Figure 3.3: Without unique IP addresses, unambiguous global routing is impossible. If the PC requests a web page from 10.1.1.2, which server will it reach?

Subnets

By applying a ***subnet mask*** (also called a ***network mask***, or simply ***netmask***) to an IP address, you can logically define both a host and the network to which it belongs. Traditionally, subnet masks are expressed using dotted decimal form, much like an IP address. For example, 255.255.255.0 is one common netmask. You will find this notation used when configuring network interfaces, creating routes, etc. However, subnet masks are more succinctly expressed using ***CIDR notation***, which simply enumerates the number of bits in the mask after a forward slash (/). Thus, 255.255.255.0 can be simplified as /24. CIDR is short for ***Classless Inter-Domain Routing***, and is defined in RFC1518[1].

A subnet mask determines the size of a given network. Using a /24 netmask, 8 bits are reserved for hosts (32 bits total - 24 bits of netmask = 8 bits for hosts). This yields up to 256 possible host addresses ($2^8 = 256$). By convention, the first value is taken as the ***network address*** (.0 or 00000000), and the last value is taken as the ***broadcast address*** (.255 or 11111111). This leaves 254 addresses available for hosts on this network.

Subnet masks work by applying AND logic to the 32 bit IP number. In binary notation, the "1" bits in the mask indicate the network address portion, and "0" bits indicate the host address portion. A logical AND is performed by comparing two bits. The result is "1" if both of the bits being compared are

---

1. RFC is short for Request For Comments. RFCs are a numbered series of documents published by the Internet Society that document ideas and concepts related to Internet technologies. Not all RFCs are actual standards. RFCs can be viewed online at http://rfc.net/

also "1". Otherwise the result is "0". Here are all of the possible outcomes of a binary AND comparison between two bits.

Bit 1	Bit 2	Result
0	0	0
0	1	0
1	0	0
1	1	1

To understand how a netmask is applied to an IP address, first convert everything to binary. The netmask 255.255.255.0 in binary contains twenty-four "1" bits:

```
255      255      255      0
11111111.11111111.11111111.00000000
```

When this netmask is combined with the IP address 10.10.10.10, we can apply a logical AND to each of the bits to determine the network address.

```
  10.10.10.10: 00001010.00001010.00001010.00001010
255.255.255.0: 11111111.11111111.11111111.00000000
        ---------------------------------
   10.10.10.0: 00001010.00001010.00001010.00000000
```

This results in the network 10.10.10.0/24. This network consists of the hosts 10.10.10.1 through 10.10.10.254, with 10.10.10.0 as the network address and 10.10.10.255 as the broadcast address.

Subnet masks are not limited to entire octets. One can also specify subnet masks like 255.254.0.0 (or /15 CIDR). This is a large block, containing 131,072 addresses, from 10.0.0.0 to 10.1.255.255. It could be further subdivided, for example into 512 subnets of 256 addresses each. The first one would be 10.0.0.0-10.0.0.255, then 10.0.1.0-10.0.1.255, and so on up to 10.1.255.0-10.1.255.255. Alternatively, it could be subdivided into 2 blocks of 65,536 addresses, or 8192 blocks of 16 addresses, or in many other ways. It could even be subdivided into a mixture of different block sizes, as long as none of them overlap, and each is a valid subnet whose size is a power of two.

While many netmasks are possible, common netmasks include:

CIDR	Decimal	# of Hosts
/30	255.255.255.252	4
/29	255.255.255.248	8
/28	255.255.255.240	16
/27	255.255.255.224	32
/26	255.255.255.192	64
/25	255.255.255.128	128
/24	255.255.255.0	256
/16	255.255.0.0	65 536
/8	255.0.0.0	16 777 216

With each reduction in the CIDR value the IP space is doubled. Remember that two IP addresses within each network are always reserved for the network and broadcast addresses.

There are three common netmasks that have special names. A /8 network (with a netmask of 255.0.0.0) defines a ***Class A*** network. A /16 (255.255.0.0) is a ***Class B***, and a /24 (255.255.255.0) is called a ***Class C***. These names were around long before CIDR notation, but are still often used for historical reasons.

Global IP Addresses

Have you ever wondered who controls the allocation of IP space? ***Globally routable IP addresses*** are assigned and distributed by ***Regional Internet Registrars*** (***RIR***s) to ISPs. The ISP then allocates smaller IP blocks to their clients as required. Virtually all Internet users obtain their IP addresses from an ISP.

The 4 billion available IP addresses are administered by the ***Internet Assigned Numbers Authority*** (***IANA***, *http://www.iana.org/*). IANA has divided this space into large subnets, usually /8 subnets with 16 million addresses each. These subnets are delegated to one of the five regional Internet registries (RIRs), which are given authority over large geographic areas.

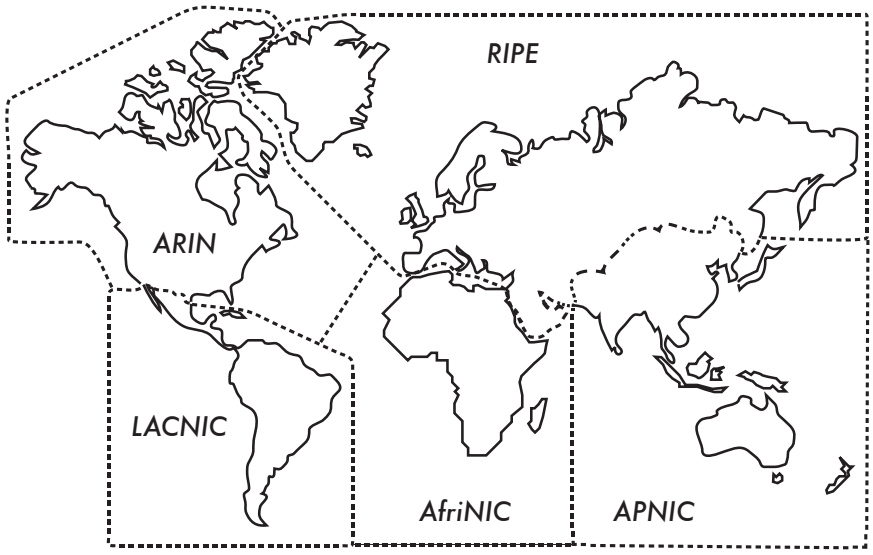


Figure 3.4: Authority for Internet IP address assignments is delegated to the five Regional Internet Registrars.

The five RIRs are:

- African Network Information Centre (AfriNIC, *http://www.afrinic.net/*)
- Asia Pacific Network Information Centre (APNIC, *http://www.apnic.net/*)
- American Registry for Internet Numbers (ARIN, *http://www.arin.net/*)
- Regional Latin-American and Caribbean IP Address Registry (LACNIC, *http://www.lacnic.net/*)
- Réseaux IP Européens (RIPE NCC, *http://www.ripe.net/*)

Your ISP will assign globally routable IP address space to you from the pool allocated to it by your RIR. The registry system assures that IP addresses are not reused in any part of the network anywhere in the world.

Once IP address assignments have been agreed upon, it is possible to pass packets between networks and participate in the global Internet. The process of moving packets between networks is called ***routing***.

Static IP Addresses

A static IP address is an address assignment that never changes. Static IP addresses are important because servers using these addresses may have DNS mappings pointed towards them, and typically serve information to other machines (such as email services, web servers, etc.).

Blocks of static IP addresses may be assigned by your ISP, either by request or automatically depending on your means of connection to the Internet.

Dynamic IP Addresses

Dynamic IP addresses are assigned by an ISP for non-permanent nodes connecting to the Internet, such as a home computer which is on a dial-up connection.

Dynamic IP addresses can be assigned automatically using the ***Dynamic Host Configuration Protocol*** (***DHCP***), or the ***Point-to-Point Protocol*** (***PPP***), depending on the type of Internet connection. A node using DHCP first requests an IP address assignment from the network, and automatically configures its network interface. IP addresses can be assigned randomly from a pool by your ISP, or might be assigned according to a policy. IP addresses assigned by DHCP are valid for a specified time (called the ***lease time***). The node must renew the DHCP lease before the lease time expires. Upon renewal, the node may receive the same IP address or a different one from the pool of available addresses.

Dynamic addresses are popular with Internet service providers, because it enables them to use fewer IP addresses than their total number of customers. They only need an address for each customer who is **active at any one time**. Globally routable IP addresses cost money, and some authorities that specialize in the assignment of addresses (such as RIPE, the European RIR) are very strict on IP address usage for ISP's. Assigning addresses dynamically allows ISPs to save money, and they will often charge extra to provide a static IP address to their customers.

Private IP addresses

Most private networks do not require the allocation of globally routable, public IP addresses for every computer in the organization. In particular, computers which are not public servers do not need to be addressable from the public Internet. Organizations typically use IP addresses from the ***private address space*** for machines on the internal network.

There are currently three blocks of private address space reserved by IANA: 10.0.0.0/8, 172.16.0.0/12, and 192.168.0.0/16. These are defined in RFC1918. These addresses are not intended to be routed on the Internet, and are typically unique only within an organization or group of organizations which choose to follow the same numbering scheme.

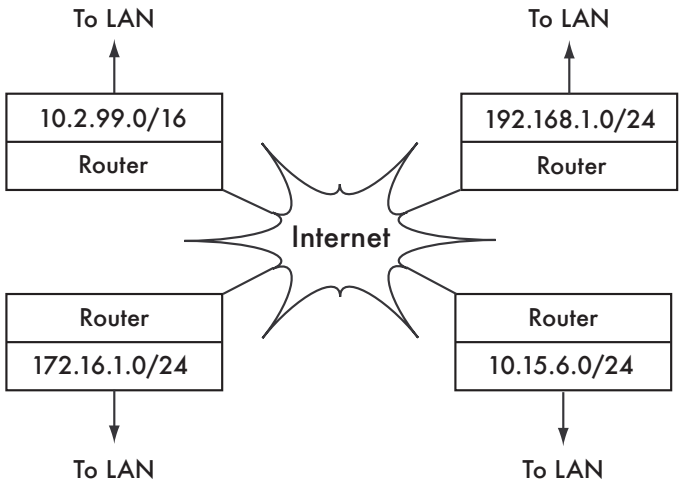


Figure 3.5: RFC1918 private addresses may be used within an organization, and are not routed on the global Internet.

If you ever intend to link together private networks that use RFC1918 address space, be sure to use unique addresses throughout all of the networks. For example, you might break the 10.0.0.0/8 address space into multiple Class B networks (10.1.0.0/16, 10.2.0.0/16, etc.). One block could be assigned to each network according to its physical location (the campus main branch, field office one, field office two, dormitories, and so forth). The network administrators at each location can then break the network down further into multiple Class C networks (10.1.1.0/24, 10.1.2.0/24, etc.) or into blocks of any other logical size. In the future, should the networks ever be linked (either by a physical connection, wireless link, or VPN), then all of the machines will be reachable from any point in the network without having to renumber network devices.

Some Internet providers may allocate private addresses like these instead of public addresses to their customers, although this has serious disadvantages. Since these addresses cannot be routed over the Internet, computers which use them are not really "part" of the Internet, and are not directly reachable from it. In order to allow them to communicate with the Internet, their private addresses must be translated to public addresses. This translation process is known as ***Network Address Translation*** (***NAT***), and is normally performed at the gateway between the private network and the Internet. We will look at NAT in more detail on **Page 43**.

Routing

Imagine a network with three hosts: A, B, and C. They use the corresponding IP addresses 192.168.1.1, 192.168.1.2 and 192.168.1.3. These hosts are part of a /24 network (their network mask is 255.255.255.0).

For two hosts to communicate on a local network, they must determine each others' MAC addresses. It is possible to manually configure each host with a mapping table from IP address to MAC address, but normally the ***Address Resolution Protocol*** (***ARP***) is used to determine this automatically.

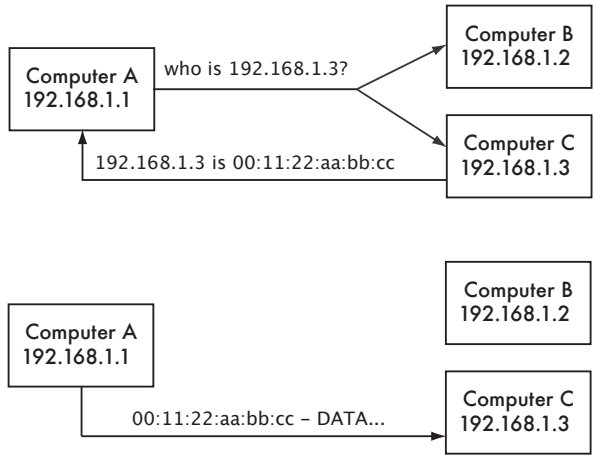


Figure 3.6: Computer A needs to send data to 192.168.1.3. But it must first ask the whole network for the MAC address that responds to 192.168.1.3.

When using ARP, host A broadcasts to all hosts the question, "Who has the MAC address for the IP 192.168.1.3?" When host C sees an ARP request for its own IP address, it replies with its MAC address.

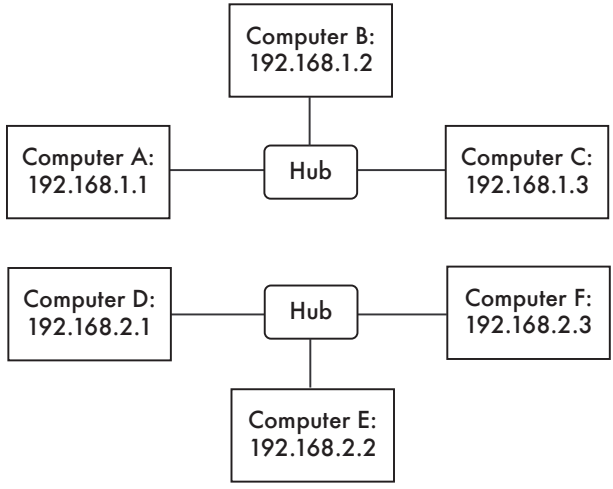


Figure 3.7: Two separate IP networks.

Consider now another network with 3 hosts, D, E, and F, with the corresponding IP addresses 192.168.2.1, 192.168.2.2, and 192.168.2.3. This is another /24 network, but it is not in the same range as the network above. All three

hosts can reach each other directly (first using ARP to resolve the IP address into a MAC address, and then sending packets to that MAC address).

Now we will add host G. This host has two network cards, with one plugged into each network. The first network card uses the IP address 192.168.1.4, and the other uses 192.168.2.4. Host G is now link-local to both networks, and can route packets between them.

But what if hosts A, B, and C want to reach hosts D, E, and F? They will need to add a route to the other network via host G. For example, hosts A-C would add a route via 192.168.1.4. In Linux, this can be accomplished with the following command:

```
# ip route add 192.168.2.0/24 via 192.168.1.4
```

...and hosts D-F would add the following:

```
# ip route add 192.168.1.0/24 via 192.168.2.4
```

The result is shown in **Figure 3.8**. Notice that the route is added via the IP address on host G that is link-local to the respective network. Host A could not add a route via 192.168.2.4, even though it is the same physical machine as 192.168.1.4 (host G), since that IP is not link-local.

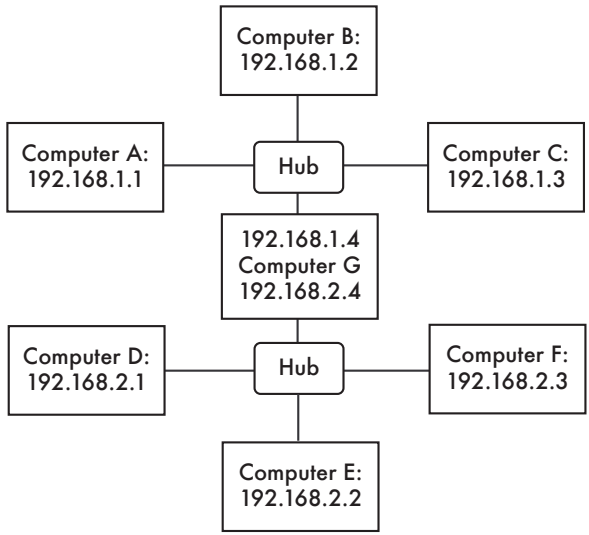


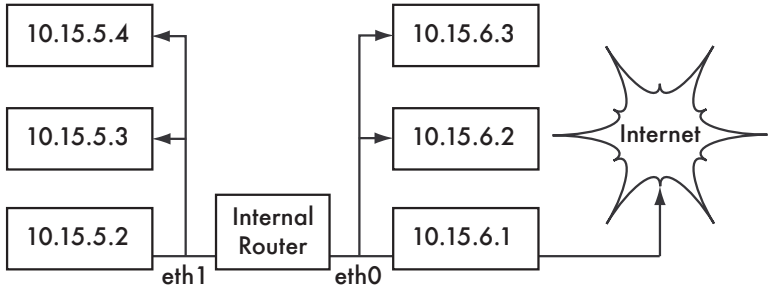
Figure 3.8: Host G acts as a router between the two networks.

A route tells the OS that the desired network doesn't lie on the immediate link-local network, and it must ***forward*** the traffic through the specified router. If host A wants to send a packet to host F, it would first send it to host G. Host G would then look up host F in its routing table, and see that it has a direct

connection to host F's network. Finally, host G would resolve the hardware (MAC) address of host F and forward the packet to it.

This is a very simple routing example, where the destination is only a single ***hop*** away from the source. As networks get more complex, many hops may need to be traversed to reach the ultimate destination. Since it isn't practical for every machine on the Internet to know the route to every other, we make use of a routing entry known as the ***default route*** (also known as the ***default gateway***). When a router receives a packet destined for a network for which it has no explicit route, the packet is forwarded to its default gateway.

The default gateway is typically the best route out of your network, usually in the direction of your ISP. An example of a router that uses a default gateway is shown in **Figure 3.9**.




```
Routing table for internal router:

Destination     Gateway     Genmask         Flags  Metric  Iface
10.15.5.0       *           255.255.255.0   U      0       eth1
10.15.6.0       *           255.255.255.0   U      0       eth0
default         10.15.6.1   0.0.0.0         UG     0       eth0
```

Figure 3.9: When no explicit route exists to a particular destination, a host uses the default gateway entry in its routing table.

Routes can be updated manually, or can dynamically react to network outages and other events. Some examples of popular dynamic routing protocols are RIP, OSPF, BGP, and OLSR. Configuring dynamic routing is beyond the scope of this book, but for further reading on the subject, see the resources in **Appendix A**.

Network Address Translation (NAT)

In order to reach hosts on the Internet, RFC1918 addresses must be converted to global, publicly routable IP addresses. This is achieved using a technique known as ***Network Address Translation***, or ***NAT***. A NAT device is a router that manipulates the addresses of packets instead of simply forwarding them. On a NAT router, the Internet connection uses one (or more) glob-

ally routed IP addresses, while the private network uses an IP address from the RFC1918 private address range. The NAT router allows the global address(es) to be shared with all of the inside users, who all use private addresses. It converts the packets from one form of addressing to the other as the packets pass through it. As far as the network users can tell, they are directly connected to the Internet and require no special software or drivers. They simply use the NAT router as their default gateway, and address packets as they normally would. The NAT router translates outbound packets to use the global IP address as they leave the network, and translates them back again as they are received from the Internet.

The major consequence of using NAT is that machines from the Internet cannot easily reach servers within the organization without setting up explicit forwarding rules on the router. Connections initiated from within the private address space generally have no trouble, although some applications (such as Voice over IP and some VPN software) can have difficulty dealing with NAT.

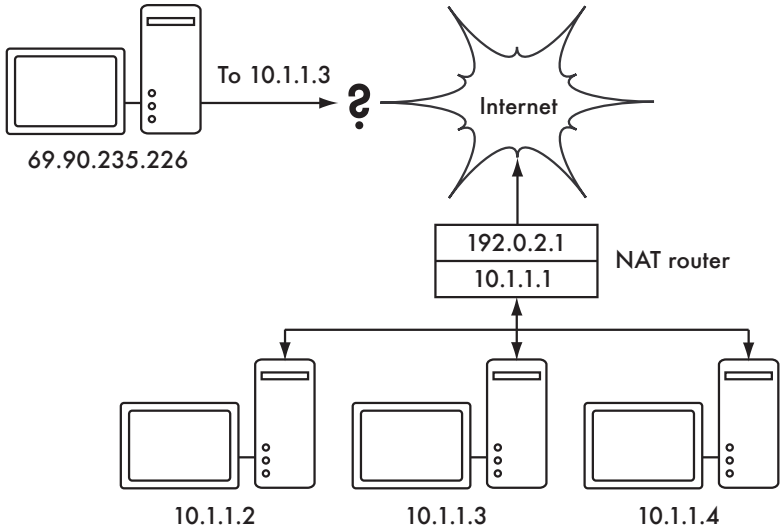


Figure 3.10: Network Address Translation allows you to share a single IP address with many internal hosts, but can make it difficult for some services to work properly.

Depending on your point of view, this can be considered a bug (since it makes it harder to set up two-way communication) or a feature (since it effectively provides a "free" firewall for your entire organization). RFC1918 addresses should be filtered on the edge of your network to prevent accidental or malicious RFC1918 traffic entering or leaving your network. While NAT performs some firewall-like functions, it is not a replacement for a real firewall.

Internet Protocol Suite

Machines on the Internet use the Internet Protocol (IP) to reach each other, even when separated by many intermediary machines. There are a number of protocols that are run in conjunction with IP that provide features as critical to normal operations as IP itself. Every packet specifies a protocol number which identifies the packet as one of these protocols. The most commonly used protocols are the ***Transmission Control Protocol*** (***TCP***, number 6), ***User Datagram Protocol*** (***UDP***, number 17), and the ***Internet Control Message Protocol*** (***ICMP***, number 1). Taken as a group, these protocols (and others) are known as the ***Internet Protocol Suite***, or simply ***TCP/IP*** for short.

The TCP and UDP protocols introduce the concept of port numbers. Port numbers allow multiple services to be run on the same IP address, and still be distinguished from each other. Every packet has a source and destination port number. Some port numbers are well defined standards, used to reach well known services such as email and web servers. For example, web servers normally ***listen*** on TCP port 80, and SMTP email servers listen on TCP port 25. When we say that a service "listens" on a port (such as port 80), we mean that it will accept packets that use its IP as the destination IP address, and 80 as the destination port. Servers usually do not care about the source IP or source port, although sometimes they will use them to establish the identity of the other side. When sending a response to such packets, the server will use its own IP as the source IP, and 80 as the source port.

When a client connects to a service, it may use any source port number on its side which is not already in use, but it must connect to the proper port on the server (e.g. 80 for web, 25 for email). TCP is a ***session oriented*** protocol with guaranteed delivery and transmission control features (such as detection and mitigation of network congestion, retries, packet reordering and reassembly, etc.). UDP is designed for ***connectionless*** streams of information, and does not guarantee delivery at all, or in any particular order.

The ICMP protocol is designed for debugging and maintenance on the Internet. Rather than port numbers, it has ***message types***, which are also numbers. Different message types are used to request a simple response from another computer (echo request), notify the sender of another packet of a possible routing loop (time exceeded), or inform the sender that a packet that could not be delivered due to firewall rules or other problems (destination unreachable).

By now you should have a solid understanding of how computers on the network are addressed, and how information flows on the network between them. Now let's take a brief look at the physical hardware that implements these network protocols.

Ethernet

Ethernet is the name of the most popular standard for connecting together computers on a ***Local Area Network*** (***LAN***). It is sometimes used to connect individual computers to the Internet, via a router, ADSL modem, or wireless device. However, if you connect a single computer to the Internet, you may not use Ethernet at all. The name comes from the physical concept of the ether, the medium which was once supposed to carry light waves through free space. The official standard is called IEEE 802.3.

The most common Ethernet standard is called 100baseT. This defines a data rate of 100 megabits per second, running over twisted pair wires, with modular RJ-45 connectors on the end. The network topology is a star, with switches or hubs at the center of each star, and end nodes (devices and additional switches) at the edges.

MAC addresses

Every device connected to an Ethernet network has a unique MAC address, assigned by the manufacturer of the network card. Its function is like that of an IP address, since it serves as a unique identifier that enables devices to talk to each other. However, the scope of a MAC address is limited to a broadcast domain, which is defined as all the computers connected together by wires, hubs, switches, and bridges, but not crossing routers or Internet gateways. MAC addresses are never used directly on the Internet, and are not transmitted across routers.

Hubs

Ethernet ***hubs*** connect multiple twisted-pair Ethernet devices together. They work at the physical layer (the lowest or first layer). They repeat the signals received by each port out to all of the other ports. Hubs can therefore be considered to be simple repeaters. Due to this design, only one port can successfully transmit at a time. If two devices transmit at the same time, they corrupt each other's transmissions, and both must back off and retransmit their packets later. This is known as a ***collision***, and each host remains responsible for detecting collisions during transmission, and retransmitting its own packets when needed.

When problems such as excessive collisions are detected on a port, some hubs can disconnect (***partition***) that port for a while to limit its impact on the rest of the network. While a port is partitioned, devices attached to it cannot communicate with the rest of the network. Hub-based networks are generally more robust than coaxial Ethernet (also known as 10base2 or ThinNet), where misbehaving devices can disable the entire segment. But hubs are limited in their usefulness, since they can easily become points of congestion on busy networks.

Switches

A ***switch*** is a device which operates much like a hub, but provides a dedicated (or ***switched***) connection between ports. Rather than repeating all traffic on every port, the switch determines which ports are communicating directly and temporarily connects them together. Switches generally provide much better performance than hubs, especially on busy networks with many computers. They are not much more expensive than hubs, and are replacing them in many situations.

Switches work at the data link layer (the second layer), since they interpret and act upon the MAC address in the packets they receive. When a packet arrives at a port on a switch, it makes a note of the source MAC address, which it associates with that port. It stores this information in an internal ***MAC table***. The switch then looks up the destination MAC address in its MAC table, and transmits the packet on the matching port. If the destination MAC address is not found in the MAC table, the packet is then sent to all of the connected interfaces. If the destination port matches the incoming port, the packet is filtered and is not forwarded.

Hubs vs. Switches

Hubs are considered to be fairly unsophisticated devices, since they inefficiently rebroadcast all traffic on every port. This simplicity introduces both a performance penalty and a security issue. Overall performance is slower, since the available bandwidth must be shared between all ports. Since all traffic is seen by all ports, any host on the network can easily monitor all of the network traffic.

Switches create virtual connections between receiving and transmitting ports. This yields better performance because many virtual connections can be made simultaneously. More expensive switches can switch traffic by inspecting packets at higher levels (at the transport or application layer), allow the creation of VLANs, and implement other advanced features.

A hub can be used when repetition of traffic on all ports is desirable; for example, when you want to explicitly allow a monitoring machine to see all of the traffic on the network. Most switches provide ***monitor port*** functionality that enables repeating on an assigned port specifically for this purpose.

Hubs were once cheaper than switches. However, the price of switches have reduced dramatically over the years. Therefore, old network hubs should be replaced whenever possible with new switches.

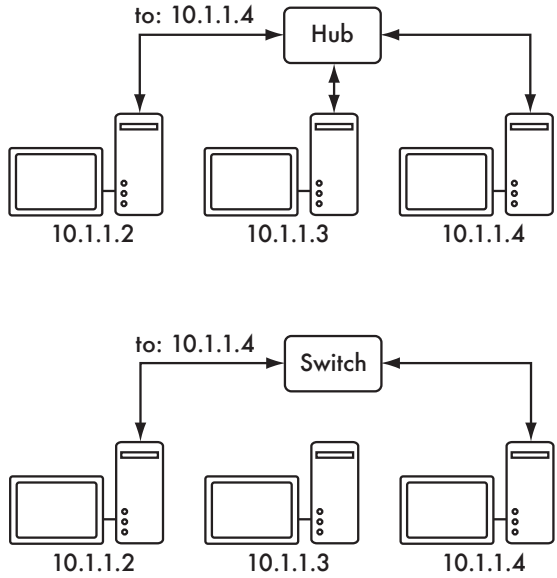


Figure 3.11: A hub simply repeats all traffic on every port, while a switch makes a temporary, dedicated connection between the ports that need to communicate.

Both hubs and switches may offer ***managed*** services. Some of these services include the ability to set the link speed (10baseT, 100baseT, 1000baseT, full or half duplex) per port, enable triggers to watch for network events (such as changes in MAC address or malformed packets), and usually include ***port counters*** for easy bandwidth accounting. A managed switch that provides upload and download byte counts for every physical port can greatly simplify network monitoring. These services are typically available via SNMP, or they may be accessed via telnet, ssh, a web interface, or a custom configuration tool.

Routers and firewalls

While hubs and switches provide connectivity on a local network segment, a router's job is to forward packets between different network segments. A router typically has two or more physical network interfaces. It may include support for different types of network media, such as Ethernet, ATM, DSL, or dial-up. Routers can be dedicated hardware devices (such as Cisco or Juniper routers) or they can be made from a standard PC with multiple network cards and appropriate software.

Routers sit at the ***edge*** of two or more networks. By definition, they have one connection to each network, and as border machines they may take on other responsibilities as well as routing. Many routers have ***firewall*** capabilities that provide a mechanism to filter or redirect packets that do not fit security or

access policy requirements. They may also provide Network Address Translation (NAT) services.

Routers vary widely in cost and capabilities. The lowest cost and least flexible are simple, dedicated hardware devices, often with NAT functionality, used to share an Internet connection between a few computers. The next step up is a software router, which consists of an operating system running on a standard PC with multiple network interfaces. Standard operating systems such as Microsoft Windows, Linux, and BSD are all capable of routing, and are much more flexible than the low-cost hardware devices. However, they suffer from the same problems as conventional PCs, with high power consumption, a large number of complex and potentially unreliable parts, and more involved configuration.

The most expensive devices are high-end dedicated hardware routers, made by companies like Cisco and Juniper. They tend to have much better performance, more features, and higher reliability than software routers on PCs. It is also possible to purchase technical support and maintenance contracts for them.

Most modern routers offer mechanisms to monitor and record performance remotely, usually via the Simple Network Management Protocol (SNMP), although the least expensive devices often omit this feature.

Other equipment

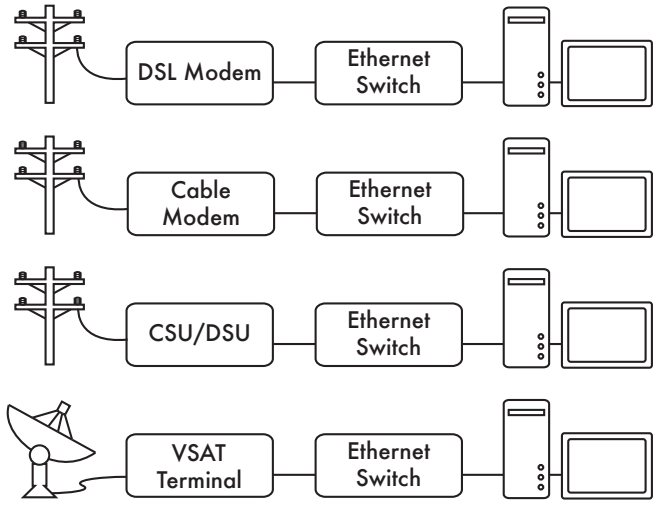


Figure 3.12: Many DSL modems, cable modems, CSU/DSUs, wireless access points, and VSAT terminals terminate at an Ethernet jack.

Each physical network has an associated piece of terminal equipment. For example, VSAT connections consist of a satellite dish connected to a termi-

nal that either plugs into a card inside a PC, or ends at a standard Ethernet connection. DSL lines use a ***DSL modem*** that bridges the telephone line to a local device, either an Ethernet network or a single computer via USB. ***Cable modems*** bridge the television cable to Ethernet, or to an internal PC card bus. Some kinds of telecom circuit (such as a T1 or T3) use a CSU/DSU to bridge the circuit to a serial port or Ethernet. Standard dialup lines use modems to connect a computer to the telephone, usually via a plug-in card or serial port. And there are many different kinds of wireless networking equipment that connect to a variety of radios and antennas, but nearly always end at an Ethernet jack.

The functionality of these devices can vary significantly between manufacturers. Some provide mechanisms for monitoring performance, while others may not. Since your Internet connection ultimately comes from your ISP, you should follow their recommendations when choosing equipment that bridges their network to your Ethernet network.

Putting it all together

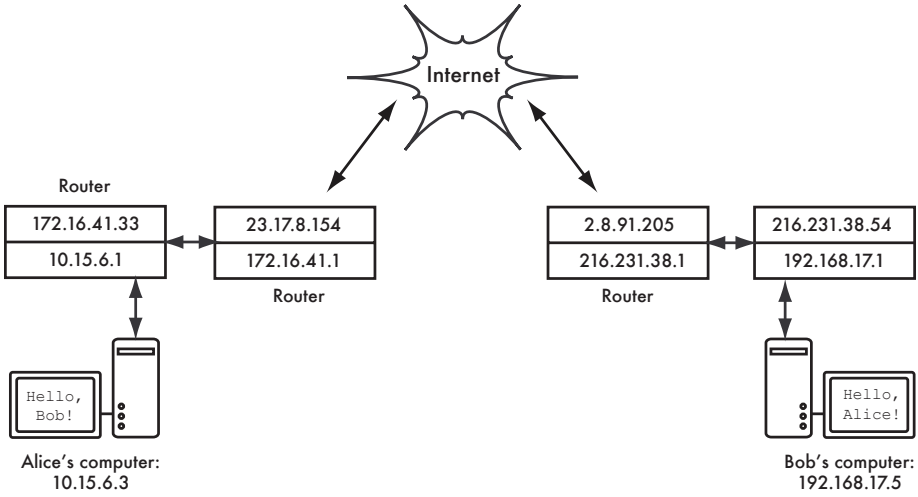


Figure 3.13: Internet networking. Each network segment has a router with two IP addresses, making it "link local" to two different networks. Packets are forwarded between routers until they reach their ultimate destination.

Once all network nodes have an IP address, they can send data packets to the IP address of any other node. Through the use of routing and forwarding, these packets can reach nodes on networks that are not physically connected to the originating node. This process describes much of what "happens" on the Internet.

In this example, you can see the path that the packets take as Alice chats with Bob using an instant messaging service. Each dotted line represents an

Ethernet cable, a wireless link, or any other kind of physical network. The cloud symbol is commonly used to stand in for "The Internet", and represents any number of intervening IP networks. Neither Alice nor Bob need to be concerned with how those networks operate, as long as the routers forward IP traffic towards the ultimate destination. If it weren't for Internet protocols and the cooperation of everyone on the net, this kind of communication would be impossible.

Designing the physical network

It may seem odd to talk about the "physical" network when building wireless networks. After all, where is the physical part of the network? In wireless networks, the physical medium we use for communication is obviously electromagnetic energy. But in the context of this chapter, the physical network refers to the mundane topic of where to put things. How do you arrange the equipment so that you can reach your wireless clients? Whether they fill an office building or stretch across many miles, wireless networks are naturally arranged in these three logical configurations: ***point-to-point links***, ***point-to-multipoint links***, and ***multipoint-to-multipoint clouds***. While different parts of your network can take advantage of all three of these configurations, any individual link will fall into one of these topologies.

Point-to-point

Point-to-point links typically provide an Internet connection where such access isn't otherwise available. One side of a point-to-point link will have an Internet connection, while the other uses the link to reach the Internet. For example, a university may have a fast frame relay or VSAT connection in the middle of campus, but cannot afford such a connection for an important building just off campus. If the main building has an unobstructed view of the remote site, a point-to-point connection can be used to link the two together. This can augment or even replace existing dial-up links. With proper antennas and clear line of sight, reliable point-to-point links in excess of thirty kilometers are possible.

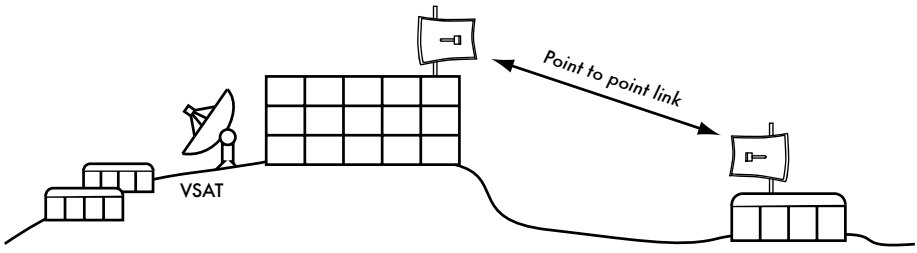


Figure 3.14: A point-to-point link allows a remote site to share a central Internet connection.

Of course, once a single point-to-point connection has been made, more can be used to extend the network even further. If the remote building in our example is at the top of a tall hill, it may be able to see other important locations that can't be seen directly from the central campus. By installing another point-to-point link at the remote site, another node can join the network and make use of the central Internet connection.

Point-to-point links don't necessarily have to involve Internet access. Suppose you have to physically drive to a remote weather monitoring station, high in the hills, in order to collect the data which it records over time. You could connect the site with a point-to-point link, allowing data collection and monitoring to happen in realtime, without the need to actually travel to the site. Wireless networks can provide enough bandwidth to carry large amounts of data (including audio and video) between any two points that have a connection to each other, even if there is no direct connection to the Internet.

Point-to-multipoint

The next most commonly encountered network layout is ***point-to-multipoint***. Whenever several nodes[2] are talking to a central point of access, this is a point-to-multipoint application. The typical example of a point-to-multipoint layout is the use of a wireless ***access point*** that provides a connection to several laptops. The laptops do not communicate with each other directly, but must be in range of the access point in order to use the network.

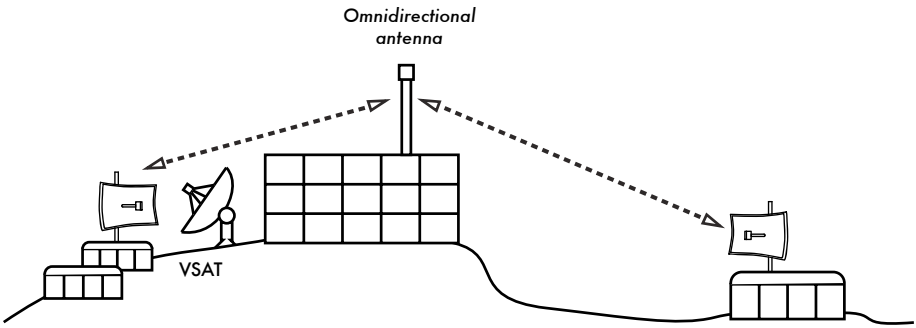


Figure 3.15: The central VSAT is now shared by multiple remote sites. All three sites can also communicate directly at speeds much faster than VSAT.

Point-to-multipoint networking can also apply to our earlier example at the university. Suppose the remote building on top of the hill is connected to the central campus with a point-to-point link. Rather than setting up several point-to-point links to distribute the Internet connection, a single antenna could be used that is visible from several remote buildings. This is a classic

---

2. A ***node*** is any device capable of sending and receiving data on a network. Access points, routers, computers, and laptops are all examples of nodes.

example of a wide area ***point*** (remote site on the hill) ***to multipoint*** (many buildings in the valley below) connection.

Note that there are a number of performance issues with using point-to-multipoint over very long distance, which will be addressed later in this chapter. Such links are possible and useful in many circumstances, but don't make the classic mistake of installing a single high powered radio tower in the middle of town and expecting to be able to serve thousands of clients, as you would with an FM radio station. As we will see, two-way data networks behave very differently than broadcast radio.

Multipoint-to-multipoint

The third type of network layout is ***multipoint-to-multipoint***, which is also referred to as an ***ad-hoc*** or ***mesh*** network. In a multipoint-to-multipoint network, there is no central authority. Every node on the network carries the traffic of every other as needed, and all nodes communicate with each other directly.

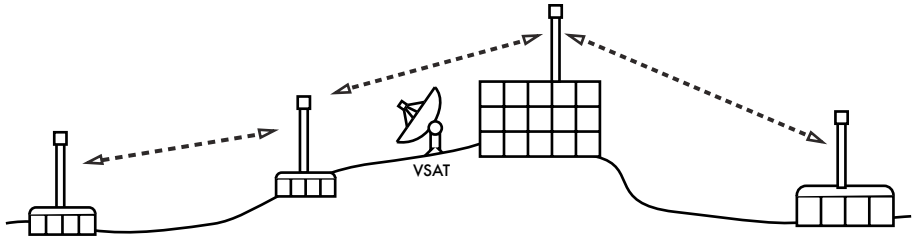


Figure 3.16: A multipoint-to-multipoint mesh. Every point can reach each other at very high speed, or use the central VSAT connection to reach the Internet.

The benefit of this network layout is that even if none of the nodes are in range of a central access point, they can still communicate with each other. Good mesh network implementations are self-healing, which means that they automatically detect routing problems and fix them as needed. Extending a mesh network is as simple as adding more nodes. If one of the nodes in the "cloud" happens to be an Internet gateway, then that connection can be shared among all of the clients.

Two big disadvantages to this topology are increased complexity and lower performance. Security in such a network is also a concern, since every participant potentially carries the traffic of every other. Multipoint-to-multipoint networks tend to be difficult to troubleshoot, due to the large number of changing variables as nodes join and leave the network. Multipoint-to-multipoint clouds typically have reduce capacity compared to point-to-point or point-to-multipoint networks, due to the additional overhead of managing the network routing and increased contention in the radio spectrum.

Nevertheless, mesh networks are useful in many circumstances. We will see an example of how to build a multipoint-to-multipoint mesh network using a routing protocol called OLSR later in this chapter.

Use the technology that fits

All of these network designs can be used to complement each other in a large network, and can obviously make use of traditional wired networking techniques whenever possible. It is a common practice, for example, to use a long distance wireless link to provide Internet access to a remote location, and then set up an access point on the remote side to provide local wireless access. One of the clients of this access point may also act as a mesh node, allowing the network to spread organically between laptop users who all ultimately use the original point-to-point link to access the Internet.

Now that we have a clear idea of how wireless networks are typically arranged, we can begin to understand how communication is possible over such networks.

802.11 wireless networks

Before packets can be forwarded and routed to the Internet, layers one (the physical) and two (the data link) need to be connected. Without link local connectivity, network nodes cannot talk to each other and route packets.

To provide physical connectivity, wireless network devices must operate in the same part of the radio spectrum. As we saw in **Chapter 2**, this means that 802.11a radios will talk to 802.11a radios at around 5 GHz, and 802.11b/g radios will talk to other 802.11b/g radios at around 2.4 GHz. But an 802.11a device cannot interoperate with an 802.11b/g device, since they use completely different parts of the electromagnetic spectrum.

More specifically, wireless cards must agree on a common channel. If one 802.11b radio card is set to channel 2 while another is set to channel 11, then the radios cannot communicate with each other.

When two wireless cards are configured to use the same protocol on the same radio channel, then they are ready to negotiate data link layer connectivity. Each 802.11a/b/g device can operate in one of four possible modes:

1. ***Master mode*** (also called ***AP*** or ***infrastructure mode***) is used to create a service that looks like a traditional access point. The wireless card creates a network with a specified name (called the ***SSID***) and channel, and offers network services on it. While in master mode, wireless cards manage all communications related to the network (authenticating wire-

less clients, handling channel contention, repeating packets, etc.) Wireless cards in master mode can only communicate with cards that are associated with it in managed mode.

2. ***Managed mode*** is sometimes also referred to as ***client*** mode. Wireless cards in managed mode will join a network created by a master, and will automatically change their channel to match it. They then present any necessary credentials to the master, and if those credentials are accepted, they are said to be ***associated*** with the master. Managed mode cards do not communicate with each other directly, and will only communicate with an associated master.

3. ***Ad-hoc mode*** creates a multipoint-to-multipoint network where there is no single master node or AP. In ad-hoc mode, each wireless card communicates directly with its neighbors. Nodes must be in range of each other to communicate, and must agree on a network name and channel.

4. ***Monitor mode*** is used by some tools (such as **Kismet**, see **Chapter 6**) to passively listen to all radio traffic on a given channel. When in monitor mode, wireless cards transmit no data. This is useful for analyzing problems on a wireless link or observing spectrum usage in the local area. Monitor mode is not used for normal communications.

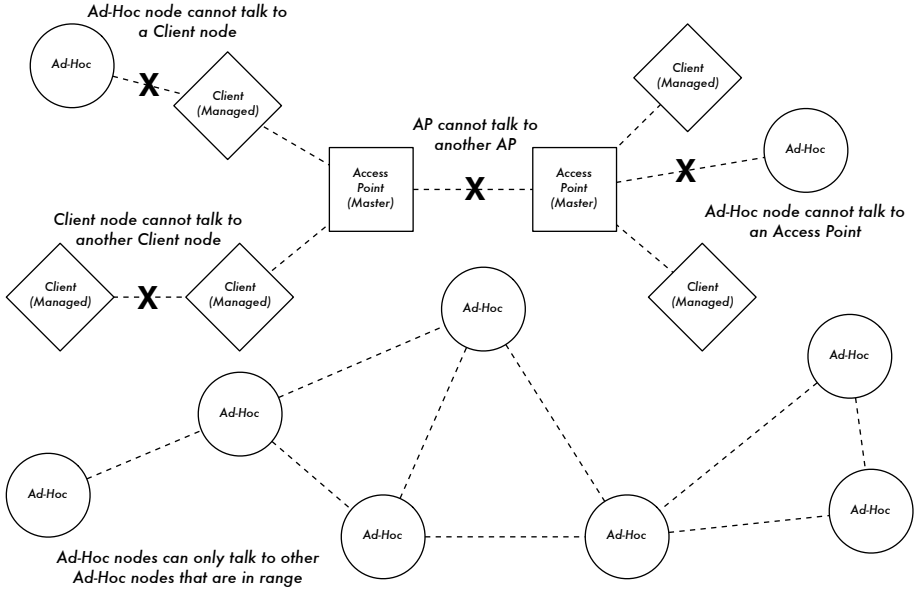


Figure 3.17: APs, Clients, and Ad-Hoc nodes.

When implementing a point-to-point or point-to-multipoint link, one radio will typically operate in master mode, while the other(s) operate in managed mode. In a multipoint-to-multipoint mesh, the radios all operate in ad-hoc mode so that they can communicate with each other directly.

It is important to keep these modes in mind when designing your network layout. Remember that managed mode clients cannot communicate with each other directly, so it is likely that you will want to run a high repeater site in master or ad-hoc mode. As we will see later in this chapter, ad-hoc is more flexible but has a number of performance issues as compared to using the master / managed modes.

Mesh networking with OLSR

Most WiFi networks operate in infrastructure mode - they consist of an access point somewhere (with a radio operating in master mode), attached to a DSL line or other large scale wired network. In such a ***hotspot*** the access point usually acts as a master station that is distributing Internet access to its clients, which operate in managed mode. This topology is similar to a mobile phone (GSM) service. Mobile phones connect to a base station - without the presence of such a base station mobiles can't communicate with each other. If you make a joke call to a friend that is sitting on the other side of the table, your phone sends data to the base station of your provider that may be a mile away - the base station then sends data back to the phone of your friend.

WiFi cards in managed mode can't communicate directly, either. Clients - for example, two laptops on the same table - have to use the access point as a relay. Any traffic between clients connected to an access point has to be sent twice. If client A and C communicate, client A sends data to the access point B, and then the access point will retransmit the data to client C. A single transmission may have a speed of 600 kByte/sec (thats about the maximum speed you could achieve with 802.11b) in our example - thus, because the data has to be repeated by the access point before it reaches its target, the effective speed between both clients will be only 300 kByte/sec.

In ad-hoc mode there is no hierarchical master-client relationship. Nodes can communicate directly as long as they are within the range of their wireless interfaces. Thus, in our example both computers could achieve full speed when operating ad-hoc, under ideal circumstances.

The disadvantage to ad-hoc mode is that clients do not repeat traffic destined for other clients. In the access point example, if two clients A and C can't directly “see” each other with their wireless interfaces, they still can communicate as long as the AP is in the wireless range of both clients.

Ad-hoc nodes do not repeat by default, but they can effectively do the same if ***routing*** is applied. Mesh networks are based on the strategy that every mesh-enabled node acts as a relay to extend coverage of the wireless network. The more nodes, the better the radio coverage and range of the mesh cloud.

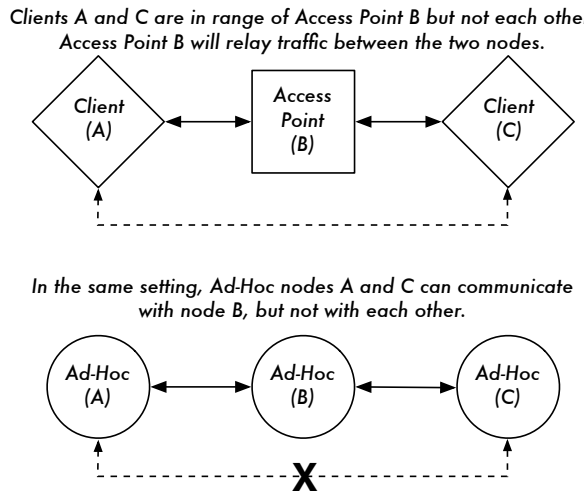


Figure 3.18: Access point B will relay traffic between clients A and C. In Ad-Hoc mode, node B will not relay traffic between A and C by default.

There is one big tradeoff that must be mentioned at this point. If the device only uses one radio interface, the available bandwidth is significantly reduced every time traffic is repeated by intermediate nodes on the way from A to B. Also, there will be interference in transmission due to nodes sharing the same channel. Thus, cheap ad-hoc mesh networks can provide good radio coverage on the last mile(s) of a community wireless network at the cost of speed-- especially if the density of nodes and transmit power is high.

If an ad-hoc network consists of only a few nodes that are up and running at all time, don't move and always have stable radio links - a long list of ifs - it is possible to write individual routing tables for all nodes by hand.

Unfortunately, those conditions are rarely met in the real world. Nodes can fail, WiFi enabled devices roam around, and interference can make radio links unusable at any time. And no one wants to update several routing tables by hand if one node is added to the network. By using routing protocols that automatically maintain individual routing tables in all nodes involved, we can avoid these issues. Popular routing protocols from the wired world (such as OSPF) do not work well in such an environment because they are not designed to deal with lossy links or rapidly changing topology.

Mesh routing with olsrd

The Optimized Link State Routing Daemon - olsrd - from *olsr.org* is a routing application developed for routing in wireless networks. We will concentrate on this routing software for several reasons. It is a open-source project that supports Mac OS X, Windows 98, 2000, XP, Linux, FreeBSD, OpenBSD and

NetBSD. Olsrd is available for access points that run Linux like the Linksys WRT54G, Asus Wl500g, AccessCube or Pocket PCs running Familiar Linux, and ships standard on Metrix kits running Pyramid. Olsrd can handle multiple interfaces and is extensible with plug-ins. It supports IPv6 and it is actively developed and used by community networks all over the world.

Note that there are several implementations of Optimized Link State Routing, which began as an IETF-draft written at INRIA France. The implementation from *olsr.org* started as a master thesis of Andreas Toennesen at UniK University. Based on practical experience of the free networking community, the routing daemon was modified. Olsrd now differs significantly from the original draft because it includes a mechanism called Link Quality Extension that measures the packet loss between nodes and calculates routes according to this information. This extension breaks compatibility to routing daemons that follow the INRIA draft. The olsrd available from *olsr.org* can be configured to behave according to the IETF draft that lacks this feature - but there is no reason to disable Link Quality Extension unless compliance with other implementations is required.

Theory

After olsrd is running for a while, a node knows about the existence of every other node in the mesh cloud and which nodes may be used to route traffic to them. Each node maintains a routing table covering the whole mesh cloud. This approach to mesh routing is called ***proactive routing***. In contrast, ***reactive routing*** algorithms seek routes only when it is necessary to send data to a specific node.

There are pros and cons to proactive routing, and there are many other ideas about how to do mesh routing that may be worth mentioning. The biggest advantage of proactive routing is that you know who is out there and you don't have to wait until a route is found. Higher protocol traffic overhead and more CPU load are among the disadvantages. In Berlin, the Freifunk community is operating a mesh cloud where olsrd has to manage more than 100 interfaces. The average CPU load caused by olsrd on a Linksys WRT54G running at 200 MHz is about 30% in the Berlin mesh. There is clearly a limit to what extent a proactive protocol can scale - depending on how many interfaces are involved and how often the routing tables are updated. Maintaining routes in a mesh cloud with static nodes takes less effort than a mesh with nodes that are constantly in motion, since the routing table has to be updated less often.

Mechanism

A node running olsrd is constantly broadcasting 'Hello' messages at a given interval so neighbors can detect it's presence. Every node computes a statistic how many 'Hellos' have been lost or received from each neighbor -

thereby gaining information about the topology and link quality of nodes in the neighborhood. The gained topology information is broadcasted as topology control messages (TC messages) and forwarded by neighbors that olsrd has chosen to be multipoint relays.

The concept of multipoint relays is a new idea in proactive routing that came up with the OLSR draft. If every node rebroadcasts topology information that it has received, unnecessary overhead can be generated. Such transmissions are redundant if a node has many neighbors. Thus, an olsrd node decides which neighbors are favorable multipoint relays that should forward its topology control messages. Note that multipoint relays are only chosen for the purpose of forwarding TC messages. Payload is routed considering all available nodes.

Two other message types exist in OLSR that announce information: whether a node offers a gateway to other networks (HNA messages) or has multiple interfaces (MID messages). There is not much to say about what this messages do apart from the fact that they exist. HNA messages make olsrd very convenient when connecting to the Internet with a mobile device. When a mesh node roams around it will detect gateways into other networks and always choose the gateway that it has the best route to. However, olsrd is by no means bullet proof. If a node announces that it is an Internet gateway - which it isn't because it never was or it is just offline at the moment - the other nodes will nevertheless trust this information. The pseudo-gateway is a black hole. To overcome this problem, a dynamic gateway plugin was written. The plugin will automatically detect at the gateway if it is actually connected and whether the link is still up. If not, olsrd ceases to send false HNA messages. It is highly recommended to build and use this plugin instead of statically enabling HNA messages.

Practice

Olsrd implements IP-based routing in a userland application - installation is pretty easy. Installation packages are available for OpenWRT, AccessCube, Mac OS X, Debian GNU/Linux and Windows. OLSR is a standard part of Metrix Pyramid. If you have to compile from source, please read the documentation that is shipped with the source package. If everything is configured properly all you have to do is start the olsr program.

First of all, it must be ensured that every node has a unique statically assigned IP-Address for each interface used for the mesh. It is not recommended (nor practicable) to use DHCP in an IP-based mesh network. A DHCP request will not be answered by a DHCP server if the node requesting DHCP needs a multihop link to connect to it, and applying dhcp relay throughout a mesh is likely impractical. This problem could be solved by using IPv6, since there is plenty of space available to generate a unique IP from the MAC address of each card involved (as suggested in "IPv6 State-

less Address Autoconfiguration in large mobile ad hoc networks" by K. Weniger and M. Zitterbart, 2002).

A wiki-page where every interested person can choose an individual IPv4 address for each interface the olsr daemon is running on may serve the purpose quite well. There is just not an easy way to automate the process if IPv4 is used.

The broadcast address should be 255.255.255.255 on mesh interfaces in general as a convention. There is no reason to enter the broadcast address explicitly, since olsrd can be configured to override the broadcast addresses with this default. It just has to be ensured that settings are the same everywhere. Olsrd can do this on its own. When a default olsrd configuration file is issued, this feature should be enabled to avoid confusion of the kind "why can't the other nodes see my machine?!?"

Now configure the wireless interface. Here is an example command how to configure a WiFi card with the name wlan0 using Linux:

```
iwconfig wlan0 essid olsr.org mode ad-hoc channel 10 rts 250 frag 256
```

Verify that the wireless part of the WiFi card has been configured so it has an ad-hoc connection to other mesh nodes within direct (single hop) range. Make sure the interface joins the same wireless channel, uses the same wireless network name ESSID (Extended Service Set IDentifier) and has the same Cell-ID as all other WiFi-Cards that build the mesh. Many WiFi cards or their respective drivers do not comply with the 802.11 standard for ad-hoc networking and may fail miserably to connect to a cell. They may be unable to connect to other devices on the same table, even if they are set up with the correct channel and wireless network name. They may even confuse other cards that behave according to the standard by creating their own Cell-ID on the same channel with the same wireless network name. WiFi cards made by Intel that are shipped with Centrino Notebooks are notorious for doing this.

You can check this out with the command **`iwconfig`** when using GNU-Linux. Here is the output on my machine:

```
wlan0 IEEE 802.11b  ESSID:"olsr.org"
 Mode:Ad-Hoc  Frequency:2.457 GHz  Cell: 02:00:81:1E:48:10
 Bit Rate:2 Mb/s   Sensitivity=1/3
 Retry min limit:8   RTS thr=250 B   Fragment thr=256 B
 Encryption key:off
 Power Management:off
 Link Quality=1/70  Signal level=-92 dBm  Noise level=-100 dBm
 Rx invalid nwid:0  Rx invalid crypt:28  Rx invalid frag:0
 Tx excessive retries:98024 Invalid misc:117503 Missed beacon:0
```

It is important to set the 'Request To Send' threshold value RTS for a mesh. There will be collisions on the radio channel between the transmissions of

nodes on the same wireless channel, and RTS will mitigate this. RTS/CTS adds a handshake before each packet transmission to make sure that the channel is clear. This adds overhead, but increases performance in case of hidden nodes - and hidden nodes are the default in a mesh! This parameter sets the size of the smallest packet (in bytes) for which the node sends RTS. The RTS threshold value must be smaller than the IP-Packet size and the 'Fragmentation threshold' value - here set to 256 - otherwise it will be disabled. TCP is very sensitive to collisions, so it is important to switch RTS on.

Fragmentation allows to split an IP packet in a burst of smaller fragments transmitted on the medium. This adds overhead, but in a noisy environment this reduces the error penalty and allows packets to get through interference bursts. Mesh networks are very noisy because nodes use the same channel and therefore transmissions are likely to interfere with each other. This parameter sets the maximum size before a data packet is split and sent in a burst - a value equal to the maximum IP packet size disables the mechanism, so it must be smaller than the IP packet size. Setting fragmentation threshold is recommended.

Once a valid IP-address and netmask is assigned and the wireless interface is up, the configuration file of olsrd must be altered in order that olsrd finds and uses the interfaces it is meant to work on.

For Mac OS-X and Windows there are nice GUI's for configuration and monitoring of the daemon available. Unfortunately this tempts users that lack background knowledge to do stupid things - like announcing black holes. On BSD and Linux the configuration file **`/etc/olsrd.conf`** has to be edited with a text editor.

A simple olsrd.conf

It is not practical to provide a complete configuration file here. These are some essential settings that should be checked.

```
UseHysteresis           no
TcRedundancy            2
MprCoverage             3
LinkQualityLevel        2
LinkQualityWinSize      20

LoadPlugin "olsrd_dyn_gw.so.0.3"
{
     PlParam     "Interval"   "60"
     PlParam     "Ping"       "151.1.1.1"
    PlParam     "Ping"       "194.25.2.129"
}

Interface "ath0" "wlan0" {
 Ip4Broadcast 255.255.255.255
}
```

There are many more options available in the **olsrd.conf**, but these basic options should get you started. After these steps have been done, olsrd can be started with a simple command in a terminal:

```
olsrd -d 2
```

I recommend to run it with the debugging option -d 2 when used on a workstation, especially for the first time. You can see what olsrd does and monitor how well the links to your neighbors are. On embedded devices the debug level should be 0 (off), because debugging creates a lot of CPU load.

The output should look something like this:

```
--- 19:27:45.51 ------------------------------------------- DIJKSTRA

192.168.120.1:1.00 (one-hop)
192.168.120.3:1.00 (one-hop)

--- 19:27:45.51 ------------------------------------------------ LINKS

IP address       hyst   LQ     lost   total  NLQ    ETX
192.168.120.1    0.000  1.000  0      20     1.000  1.00
192.168.120.3    0.000  1.000  0      20     1.000  1.00

--- 19:27:45.51 ---------------------------------------- NEIGHBORS

IP address       LQ     NLQ    SYM   MPR   MPRS  will
192.168.120.1    1.000  1.000  YES   NO    YES   3
192.168.120.3    1.000  1.000  YES   NO    YES   6

--- 19:27:45.51 ---------------------------------------- TOPOLOGY

Source IP addr   Dest IP addr     LQ     ILQ    ETX
192.168.120.1    192.168.120.17   1.000  1.000  1.00
192.168.120.3    192.168.120.17   1.000  1.000  1.00
```

Using OLSR on Ethernet and multiple interfaces

It is not necessary to have a wireless interface to test or use olsrd - although that is what olsrd is designed for. It may as well be used on any NIC. WiFi-interfaces don't have to operate always in ad-hoc mode to form a mesh when mesh nodes have more than one interface. For dedicated links it may be a very good option to have them running in infrastructure mode. Many WiFi cards and drivers are buggy in ad-hoc mode, but infrastructure mode works fine - because everybody expects at least this feature to work. Ad-hoc mode has not had many users so far, so the implementation of the ad-hoc mode was done sloppily by many manufacturers. With the rising popularity of mesh networks, the driver situation is improving now.

Many people use olsrd on wired and wireless interfaces - they don't think about network architecture. They just connect antennas to their WiFi cards, connect cables to their Ethernet cards, enable olsrd to run on all computers and all interfaces and fire it up. That is quite an abuse of a protocol that was designed to do wireless networking on lossy links - but - why not?

They expect olsrd to solve every networking problem. Clearly it is not necessary to send 'Hello' messages on a wired interface every two seconds - but it works. This should not be taken as a recommendation - it is just amazing what people do with such a protocol and that they have such success with it. In fact the idea of having a protocol that does everything for newbies that want to have a small to medium sized routed LAN is very appealing.

Plugins

A number of plugins are available for olsrd. Check out the *olsr.org* website for a complete list. Here a little HOWTO for the network topology visualization plugin **`olsrd_dot_draw`**.

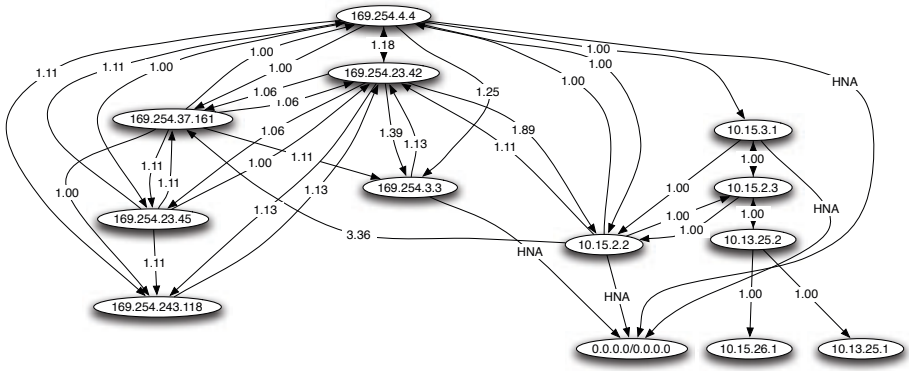


Figure 3.19: An automatically generated OLSR network topology.

Often it is very good for the understanding of a mesh network to have the ability to show the network topology graphically. **`olsrd_dot_draw`** outputs the topology in the dot file format on TCP port 2004. The graphviz tools can then be used to draw the graphs.

Installing the dot_draw Plugin

Compile the olsr plugins separately and install them. To load the plugin add the following lines to **`/etc/olsrd.conf`**. The parameter "accept" specifies which host is accepted to view the Topology Information (currently only one) and is "localhost" by default. The parameter "port" specifies the TCP port.

```
LoadPlugin "olsrd_dot_draw.so.0.3"
{
       PlParam "accept" "192.168.0.5"
       PlParam "port" "2004"
}
```

Then restart olsr and check if you get output on TCP Port 2004

```
telnet localhost 2004
```

After a while you should get some text output.

Now you can save the output graph descriptions and run the tools **`dot`** or **`neato`** form the graphviz package to get images.

Bruno Randolf has written a small perl script which continuously gets the topology information from olsrd and displays it using the graphviz and ImageMagick tools.

First install the following packages on your workstation:

- graphviz, *http://www.graphviz.org/*
- ImageMagick, *http://www.imagemagick.org/*

Download the script at: *http://meshcube.org/nylon/utils/olsr-topology-view.pl*

Now you can start the script with **`./olsr-topology-view.pl`** and view the topology updates in near-realtime.

Troubleshooting

As long as the WiFi-cards can 'see' each other directly with their radios, doing a ping will work whether olsrd is running or not. This works because the large netmasks effectively make every node link-local, so routing issues are sidestepped at the first hop. This should be checked first if things do not seem to work as expected. Most headaches people face with WiFi in Ad-Hoc mode are caused by the fact that the ad-hoc mode in drivers and cards are implemented sloppily. If it is not possible to ping nodes directly when they are in range it is most likely a card/driver issue, or your network settings are wrong.

If the machines can ping each other, but olsrd doesn't find routes, then the IP-addresses, netmask and broadcast address should be checked.

Finally, are you running a firewall? Make sure it doesn't block UDP port 698.

Estimating capacity

Wireless links can provide significantly greater ***throughput*** to users than traditional Internet connections, such as VSAT, dialup, or DSL. Throughput is also referred to as ***channel capacity***, or simply ***bandwidth*** (although this term is unrelated to radio bandwidth). It is important to understand that a wireless device's listed speed (the ***data rate***) refers to the rate at which the radios can exchange symbols, not the usable throughput you will observe. As mentioned earlier, a single 802.11g link may use 54 Mbps radios, but it will only provide up to 22 Mbps of actual throughput. The rest is overhead that the radios need in order to coordinate their signals using the 802.11g protocol.

Note that throughput is a measurement of bits over time. 22 Mbps means that in any given second, up to 22 megabits can be sent from one end of the link to the other. If users attempt to push more than 22 megabits through the link, it will take longer than one second. Since the data can't be sent immediately, it is put in a ***queue***, and transmitted as quickly as possible. This backlog of data increases the time needed for the most recently queued bits to the traverse the link. The time that it takes for data to traverse a link is called ***latency***, and high latency is commonly referred to as ***lag***. Your link will eventually send all of the queued traffic, but your users will likely complain as the lag increases.

How much throughput will your users really need? It depends on how many users you have, and how they use the wireless link. Various Internet applications require different amounts of throughput.

Application	BW / User	Notes
Text messaging / IM	< 1 kbps	As traffic is infrequent and asynchronous, IM will tolerate high latency.
Email	1 to 100 kbps	As with IM, email is asynchronous and intermittent, so it will tolerate latency. Large attachments, viruses, and spam significantly add to bandwidth usage. Note that web email services (such as Yahoo or Hotmail) should be considered as web browsing, not as email.
Web browsing	50 - 100+ kbps	Web browsers only use the network when data is requested. Communication is asynchronous, so a fair amount of lag can be tolerated. As web browsers request more data (large images, long downloads, etc.) bandwidth usage will go up significantly.

Application	BW / User	Notes
Streaming audio	96 - 160 kbps	Each user of a streaming audio service will use a constant amount of relatively large bandwidth for as long as it plays. It can tolerate some transient latency by using large buffers on the client. But extended periods of lag will cause audio "skips" or outright session failures.
Voice over IP (VoIP)	24 - 100+ kbps	As with streaming audio, VoIP commits a constant amount of bandwidth to each user for the duration of the call. But with VoIP, the bandwidth is used roughly equally in both directions. Latency on a VoIP connection is immediate and annoying to users. Lag greater than a few milliseconds is unacceptable for VoIP.
Streaming video	64 - 200+ kbps	As with streaming audio, some intermittent latency is avoided by using buffers on the client. Streaming video requires high throughput and low latency to work properly.
Peer-to-peer file-sharing applications (BitTorrent, KaZaA, Gnutella, eDonkey, etc.)	0 - infinite Mbps	While peer to peer applications will tolerate any amount of latency, they tend to use up all available throughput by transmitting data to as many clients as possible, as quickly as possible. Use of these applications will cause latency and throughput problems for all other network users unless you use careful bandwidth shaping.

To estimate the necessary throughput you will need for your network, multiply the expected number of users by the sort of application they will probably use. For example, 50 users who are chiefly browsing the web will likely consume 2.5 to 5 Mbps or more of throughput at peak times, and will tolerate some latency. On the other hand, 50 simultaneous VoIP users would require 5 Mbps or more of throughput **in both directions** with absolutely no latency. Since 802.11g wireless equipment is ***half duplex*** (that is, it only transmits or receives, never both at once) you should accordingly double the required throughput, for a total of **10 Mbps**. Your wireless links must provide that capacity every second, or conversations will lag.

Since all of your users are unlikely to use the connection at precisely the same moment, it is common practice to ***oversubscribe*** available throughput by some factor (that is, allow more users than the maximum available band-

width can support). Oversubscribing by a factor of 2 to 5 is quite common. In all likelihood, you will oversubscribe by some amount when building your network infrastructure. By carefully monitoring throughput throughout your network, you will be able to plan when to upgrade various parts of the network, and how much additional resources will be needed.

Expect that no matter how much capacity you supply, your users will eventually find applications that will use it all. As we'll see at the end of this chapter, using bandwidth shaping techniques can help mitigate some latency problems. By using bandwidth shaping, web caching, and other techniques, you can significantly reduce latency and increase overall network throughput.

To get a feeling for the lag felt on very slow connections, the ICTP has put together a bandwidth simulator. It will simultaneously download a web page at full speed and at a reduced rate that you choose. This demonstration gives you an immediate understanding of how low throughput and high latency reduce the usefulness of the Internet as a communications tool. It is available at *http://wireless.ictp.trieste.it/simulator/*

Link planning

A basic communication system consists of two radios, each with its associated antenna, the two being separated by the path to be covered. In order to have a communication between the two, the radios require a certain minimum signal to be collected by the antennas and presented to their input socket. Determining if the link is feasible is a process called ***link budget*** calculation. Whether or not signals can be passed between the radios depends on the quality of the equipment being used and on the diminishment of the signal due to distance, called ***path loss***.

Calculating the link budget

The power available in an 802.11 system can be characterized by the following factors:

- **Transmit Power**. It is expressed in milliwatts or in dBm. Transmit Power ranges from 30mW to 200mW or more. TX power is often dependent on the transmission rate. The TX power of a given device should be specified in the literature provided by the manufacturer, but can sometimes be difficult to find. Online databases such as the one provided by SeattleWireless (*http://www.seattlewireless.net/HardwareComparison*) may help.
- **Antenna Gain**. Antennas are passive devices that create the effect of amplification by virtue of their physical shape. Antennas have the same characteristics when receiving and transmitting. So a 12 dBi antenna is simply

a 12 dBi antenna, without specifying if it is in transmission or reception mode. Parabolic antennas have a gain of 19-24 dBi, omnidirectional antennas have 5-12 dBi, sectorial antennas have roughly a 12-15 dBi gain.

- **Minimum Received Signal Level,** or simply, the sensitivity of the receiver. The minimum RSL is always expressed as a negative dBm (- dBm) and is the lowest power of signal the radio can distinguish. The minimum RSL is dependent upon rate, and as a general rule the lowest rate (1 Mbps) has the greatest sensitivity. The minimum will be typically in the range of -75 to -95 dBm. Like TX power, the RSL specifications should be provided by the manufacturer of the equipment.
- **Cable Losses**. Some of the signal's energy is lost in the cables, the connectors and other devices, going from the radios to the antennas. The loss depends on the type of cable used and on its length. Signal loss for short coaxial cables including connectors is quite low, in the range of 2-3 dB. It is better to have cables as short as possible.

When calculating the path loss, several effects must be considered. One has to take into account the ***free space loss***, ***attenuation*** and ***scattering***. Signal power is diminished by geometric spreading of the wavefront, commonly known as free space loss. Ignoring everything else, the further away the two radios, the smaller the received signal is due to free space loss. This is independent from the environment, depending only on the distance. This loss happens because the radiated signal energy expands as a function of the distance from the transmitter.

Using decibels to express the loss and using 2.45 GHz as the signal frequency, the equation for the free space loss is

$$L_{fsl} = 40 + 20*\log(r)$$

where L_{fsl} is expressed in dB and r is the distance between the transmitter and receiver, in meters.

The second contribution to the path loss is given by attenuation. This takes place as some of the signal power is absorbed when the wave passes through solid objects such as trees, walls, windows and floors of buildings. Attenuation can vary greatly depending upon the structure of the object the signal is passing through, and it is very difficult to quantify. The most convenient way to express its contribution to the total loss is by adding an "allowed loss" to the free space. For example, experience shows that trees add 10 to 20 dB of loss per tree in the direct path, while walls contribute 10 to 15 dB depending upon the construction.

Along the link path, the RF energy leaves the transmitting antenna and energy spreads out. Some of the RF energy reaches the receiving antenna directly,

while some bounces off the ground. Part of the RF energy which bounces off the ground reaches the receiving antenna. Since the reflected signal has a longer way to travel, it arrives at the receiving antenna later than the direct signal. This effect is called ***multipath***, or signal dispersion. In some cases reflected signals add together and cause no problem. When they add together out of phase, the received signal is almost worthless. In some cases, the signal at the receiving antenna can be zeroed by the reflected signals. This is known as extreme fading, or ***nulling***. There is a simple technique that is used to deal with multipath, called ***antenna diversity***. It consists of adding a second antenna to the radio. Multipath is in fact a very location-specific phenomenon. If two signals add out of phase at one location, they will not add destructively at a second, nearby location. If there are two antennas, at least one of them should be able to receive a usable signal, even if the other is receiving a distorted one. In commercial devices, antenna switching diversity is used: there are multiple antennas on multiple inputs, with a single receiver. The signal is thus received through only one antenna at a time. When transmitting, the radio uses the antenna last used for reception. The distortion given by multipath degrades the ability of the receiver to recover the signal in a manner much like signal loss. A simple way of applying the effects of scattering in the calculation of the path loss is to change the exponent of the distance factor of the free space loss formula. The exponent tends to increase with the range in an environment with a lot of scattering. An exponent of 3 can be used in an outdoor environment with trees, while one of 4 can be used for an indoor environment.

When free space loss, attenuation, and scattering are combined, the path loss is:

```
L(dB) = 40 + 10*n*log(r) + L(allowed)
```

For a rough estimate of the link feasibility, one can evaluate just the free space loss. The environment can bring further signal loss, and should be considered for an exact evaluation of the link. The environment is in fact a very important factor, and should never be neglected.

To evaluate if a link is feasible, one must know the characteristics of the equipment being used and evaluate the path loss. Note that when performing this calculation, you should only add the TX power of one side of the link. If you are using different radios on either side of the link, you should calculate the path loss twice, once for each direction (using the appropriate TX power for each calculation). Adding up all the gains and subtracting all the losses gives

```
  TX Power Radio 1
+ Antenna Gain Radio 1
- Cable Losses Radio 1
+ Antenna Gain Radio 2
- Cable Losses Radio 2
_____________________________

= Total Gain
```

Subtracting the Path Loss from the Total Gain:

```
                Total Gain
              - Path Loss
              ———————————
= Signal Level at one side of the link
```

If the resulting signal level is greater than the minimum received signal level, then the link is feasible! The received signal is powerful enough for the radios to use it. Remember that the minimum RSL is always expressed as a negative dBm, so -56 dBm is greater than -70 dBm. On a given path, the variation in path loss over a period of time can be large, so a certain margin (difference between the signal level and the minimum received signal level) should be considered. This margin is the amount of signal above the sensitivity of radio that should be received in order to ensure a stable, high quality radio link during bad weather and other atmospheric disturbances. A margin of 10 to 15 dB is fine. To give some space for attenuation and multipath in the received radio signal, a margin of 20dB should be safe enough.

Once you have calculated the link budget in one direction, repeat the calculation for the other direction. Substitute the transmit power for that of the second radio, and compare the result against the minimum received signal level of the first radio.

Example link budget calculation

As an example, we want to estimate the feasibility of a 5 km link, with one access point and one client radio. The access point is connected to an omnidirectional antenna with 10 dBi gain, while the client is connected to a sectorial antenna with 14 dBi gain. The transmitting power of the AP is 100mW (or 20 dBm) and its sensitivity is -89 dBm. The transmitting power of the client is 30mW (or 15 dBm) and its sensitivity is -82 dBm. The cables are short, with a loss of 2dB at each side.

Adding up all the gains and subtracting all the losses for the AP to client link gives:

```
  20 dBm (TX Power Radio 1)
+ 10 dBi (Antenna Gain Radio 1)
-  2 dB  (Cable Losses Radio 1)
+ 14 dBi (Antenna Gain Radio 2)
-  2 dB  (Cable Losses Radio 2)
————————
  40 dB = Total Gain
```

The path loss for a 5 km link, considering only the free space loss is:

```
Path Loss = 40 + 20log(5000) = 113 dB
```

Subtracting the path loss from the total gain

```
40 dB - 113 dB = -73 dB
```

Since -73 dB is greater than the minimum receive sensitivity of the client radio (-82 dBm), the signal level is just enough for the client radio to be able to hear the access point. There is only 9 dB of margin (82 dB - 73 dB) which will likely work fine in fair weather, but may not be enough to protect against extreme weather conditions.

Next we calculate the link from the client back to the access point:

```
  15 dBm (TX Power Radio 2)
+ 14 dBi (Antenna Gain Radio 2)
-  2 dB  (Cable Losses Radio 2)
+ 10 dBi (Antenna Gain Radio 1)
-  2 dB  (Cable Losses Radio 1)
————————
  35 dB = Total Gain
```

Obviously, the path loss is the same on the return trip. So our received signal level on the access point side is:

```
35 dB - 113 dB = -78 dB
```

Since the receive sensitivity of the AP is -89dBm, this leaves us 11dB of fade margin (89dB - 78dB). Overall, this link will probably work but could use a bit more gain. By using a 24dBi dish on the client side rather than a 14dBi sectorial antenna, you will get an additional 10dBi of gain on both directions of the link (remember, antenna gain is reciprocal). A more expensive option would be to use higher power radios on both ends of the link, but note that adding an amplifier or higher powered card to one end generally does not help the overall quality of the link.

Online tools can be used to calculate the link budget. For example, the Green Bay Professional Packet Radio's Wireless Network Link Analysis (*http://my.athenet.net/~multiplx/cgi-bin/wireless.main.cgi*) is an excellent tool. The Super Edition generates a PDF file containing the Fresnel zone and radio path graphs. The calculation scripts can even be downloaded from the website and installed locally.

The Terabeam website also has excellent calculators available online (*http://www.terabeam.com/support/calculations/index.php*).

Tables for calculating link budget

To calculate the link budget, simply approximate your link distance, then fill in the following tables:

Free Space Path Loss at 2.4 GHz

Distance (m)	100	500	1,000	3,000	5,000	10,000
Loss (dB)	80	94	100	110	113	120

For more path loss distances, see **Appendix C**.

Antenna Gain:

Radio 1 Antenna	+ Radio 2 Antenna	= Total Antenna Gain

Losses:

Radio 1 + Cable Loss (dB)	Radio 2 + Cable Loss (dB)	Free Space Path Loss (dB)	= Total Loss (dB)

Link Budget for Radio 1 → Radio 2:

Radio 1 TX Power	+ Antenna Gain	- Total Loss	= Signal	> Radio 2 Sensitivity

Link Budget for Radio 2 ➔ Radio 1:

Radio 2 TX Power	+ Antenna Gain	- Total Loss	= Signal	> Radio 1 Sensitivity

If the received signal is greater than the minimum received signal strength in both directions of the link, as well as any noise received along the path, then the link is possible.

Link planning software

While calculating a link budget by hand is straightforward, there are a number of tools available that will help automate the process. In addition to calculating free space loss, these tools will take many other relevant factors into account as well (such as tree absorption, terrain effects, climate, and even estimating path loss in urban areas). In this section, we will discuss two free tools that are useful for planning wireless links: Green Bay Professional Packet Radio's online interactive network design utilities, and RadioMobile.

Interactive design CGIs

The Green Bay Professional Packet Radio group (GBPRR) has made a variety of very useful link planning tools available for free online. You can browse these tools online at *http://www.qsl.net/n9zia/wireless/page09.html .* Since the tools are available online, they will work with any device that has a web browser and Internet access.

We will look at the first tool, **Wireless Network Link Analysis**, in detail. You can find it online at *http://my.athenet.net/~multiplx/cgi-bin/wireless.main.cgi.*

To begin, enter the channel to be used on the link. This can be specified in MHz or GHz. If you don't know the frequency, consult the table in **Appendix B**. Note that the table lists the channel's center frequency, while the tool asks for the highest transmitted frequency. The difference in the ultimate result is minimal, so feel free to use the center frequency instead. To find the highest transmitted frequency for a channel, just add 11MHz to the center frequency.

Next, enter the details for the transmitter side of the link, including the transmission line type, antenna gain, and other details. Try to fill in as much data as you know or can estimate. You can also enter the antenna height and elevation for this site. This data will be used for calculating the antenna tilt

angle. For calculating Fresnel zone clearance, you will need to use GBPRR's Fresnel Zone Calculator.

The next section is very similar, but includes information about the other end of the link. Enter all available data in the appropriate fields.

Finally, the last section describes the climate, terrain, and distance of the link. Enter as much data as you know or can estimate. Link distance can be calculated by specifying the latitude and longitude of both sites, or entered by hand.

Now, click the Submit button for a detailed report about the proposed link. This includes all of the data entered, as well as the projected path loss, error rates, and uptime. These numbers are all completely theoretical, but will give you a rough idea of the feasibility of the link. By adjusting values on the form, you can play "what-if?" to see how changing various parameters will affect the connection.

In addition to the basic link analysis tool, GBPRR provides a "super edition" that will produce a PDF report, as well as a number of other very useful tools (including the Fresnel Zone Calculator, Distance & Bearing Calculator, and Decibel Conversion Calculator to name just a few). Source code to most of the tools is provided as well.

RadioMobile

Radio Mobile is a tool for the design and simulation of wireless systems. It predicts the performance of a radio link by using information about the equipment and a digital map of the area. It is public domain software that runs on Windows, or using Linux and the Wine emulator.

Radio Mobile uses a ***digital terrain elevation model*** for the calculation of coverage, indicating received signal strength at various points along the path. It automatically builds a profile between two points in the digital map showing the coverage area and first Fresnel zone. During the simulation, it checks for line of sight and calculates the Path Loss, including losses due to obstacles. It is possible to create networks of different topologies, including net master/slave, point-to-point, and point-to-multipoint. The software calculates the coverage area from the base station in a point-to-multipoint system. It works for systems having frequencies from 100 kHz to 200 GHz. ***Digital elevation maps*** (***DEM***) are available for free from several sources, and are available for most of the world. DEMs do not show coastlines or other readily identifiable landmarks, but they can easily be combined with other kinds of data (such as aerial photos or topographical charts) in several layers to obtain a more useful and readily recognizable representation. You can digitize your own maps and combine them with DEMs. The digital elevation maps can be merged with

scanned maps, satellite photos and Internet map services (such as Google Maps) to produce accurate prediction plots.

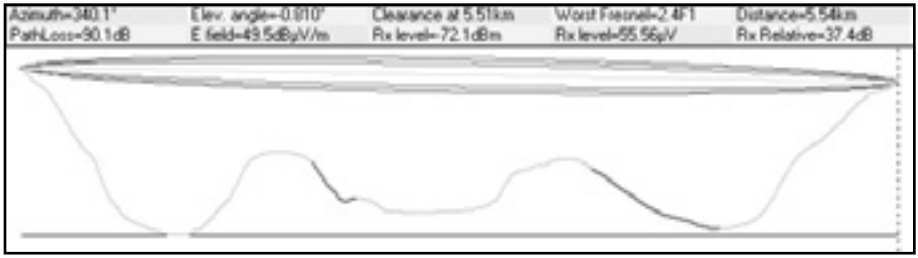


Figure 3.20: Link feasibility, including Fresnel zone and line of sight estimate, using RadioMobile.

The main Radio Mobile webpage, with examples and tutorials, is available at: *http://www.cplus.org/rmw/english1.html*

RadioMobile under Linux

Radio Mobile will also work using Wine under Ubuntu Linux. While the application runs, some button labels may run beyond the frame of the button and can be hard to read.

We were able to make Radio Mobile work with Linux using the following environment:

- IBM Thinkpad x31
- Ubuntu Breezy (v5.10), *http://www.ubuntu.com/*
- Wine version 20050725, from the Ubuntu Universe repository

There are detailed instructions for installing RadioMobile on Windows at *http://www.cplus.org/rmw/english1.html*. You should follow all of the steps except for step 1 (since it is difficult to extract a DLL from the VBRUN60SP6.EXE file under Linux). You will either need to copy the MSVBVM60.DLL file from a Windows machine that already has the Visual Basic 6 run-time environment installed, or simply Google for MSVBVM60.DLL, and download the file.

Now continue with step 2 at from the above URL, making sure to unzip the downloaded files in the same directory into which you have placed the downloaded DLL file. Note that you don't have to worry about the stuff after step 4; these are extra steps only needed for Windows users.

Finally, you can start Wine from a terminal with the command:

```
# wine RMWDLX.exe
```

You should see RadioMobile running happily in your XWindows session.

Avoiding noise

The unlicensed ISM and U-NII bands represent a very tiny piece of the known electromagnetic spectrum. Since this region can be utilized without paying license fees, many consumer devices use it for a wide range of applications. Cordless phones, analog video senders, Bluetooth, baby monitors, and even microwave ovens compete with wireless data networks for use of the very limited 2.4 GHz band. These signals, as well as other local wireless networks, can cause significant problems for long range wireless links. Here are some steps you can use to reduce reception of unwanted signals.

- **Increase antenna gain on both sides of a point-to-point link.** Antennas not only add gain to a link, but their increased directionality tends to reject noise from areas around the link. Two high gain dishes that are pointed at each other will reject noise from directions that are outside the path of the link. Using omnidirectional antennas will receive noise from all directions.

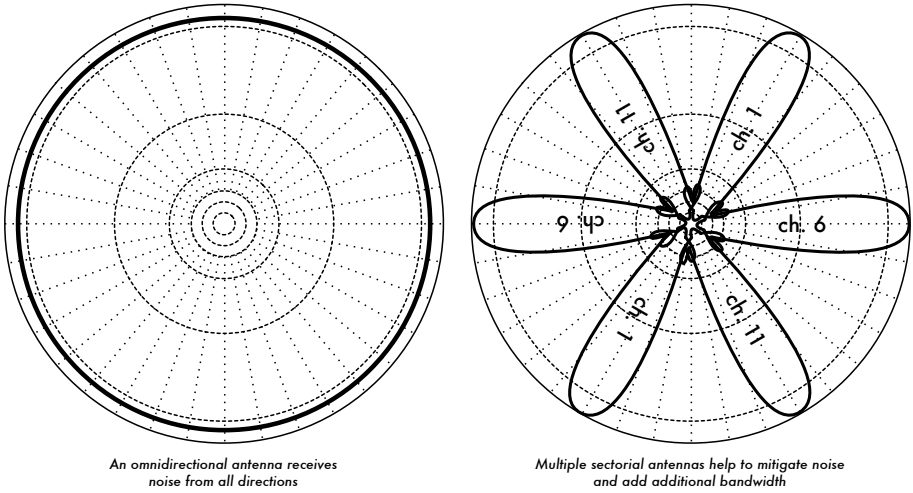


Figure 3.21: A single omnidirectional antenna vs. multiple sectorials.

- **Use sectorials instead of using an omnidirectional.** By making use of several sectorial antennas, you can reduce the overall noise received at a distribution point. By staggering the channels used on each sectorial, you can also increase the available bandwidth to your clients.

- **Don't use an amplifier.** As we will see in **Chapter 4**, amplifiers can make interference issues worse by indiscriminately amplifying all received signals, including sources of interference. Amplifiers also cause interference problems for other nearby users of the band.
- **Use the best available channel.** Remember that 802.11b/g channels are 22 MHz wide, but are only separated by 5MHz. Perform a site survey, and select a channel that is as far as possible from existing sources of interference. Remember that the wireless landscape can change at any time as people add new devices (cordless phones, other networks, etc.) If your link suddenly has trouble sending packets, you may need to perform another site survey and pick a different channel.
- **Use smaller hops and repeaters, rather than a single long distance shot.** Keep your point-to-point links as short as possible. While it may be possible to create a 12 km link that cuts across the middle of a city, you will likely have all kinds of interference problems. If you can break that link into two or three shorter hops, the link will likely be more stable. Obviously this isn't possible on long distance rural links where power and mounting structures are unavailable, but noise problems are also unlikely in those settings.
- **If possible, use 5.8 GHz, 900MHz, or another unlicensed band**. While this is only a short term solution, there is currently far more consumer equipment installed in the field that uses 2.4 GHz. Using 802.11a or a 2.4 GHz to 5.8 GHz step-up device will let you avoid this congestion altogether. If you can find it, some old 802.11 equipment uses unlicensed spectrum at 900MHz (unfortunately at much lower bit rates). Other technologies, such as Ronja (*http://ronja.twibright.com/*) use optical technology for short distance, noise-free links.
- **If all else fails, use licensed spectrum**. There are places where all available unlicensed spectrum is effectively used. In these cases, it may make sense to spend the additional money for proprietary equipment that uses a less congested band. For long distance point-to-point links that require very high throughput and maximum uptime, this is certainly an option. Of course, these features come at a much higher price tag compared to unlicensed equipment.

To identify sources of noise, you need tools that will show you what is happening in the air at 2.4 GHz. We will see some examples of these tools in **Chapter 6**.

Repeaters

The most critical component to building long distance network links is ***line of sight*** (often abbreviated as ***LOS***). Terrestrial microwave systems simply cannot tolerate large hills, trees, or other obstacles in the path of a long distance link. You must have a clear idea of the lay of the land between two points before you can determine if a link is even possible.

But even if there is a mountain between two points, remember that obstacles can sometimes be turned into assets. Mountains may block your signal, but assuming power can be provided they also make very good ***repeater*** sites.

Repeaters are nodes that are configured to rebroadcast traffic that is not destined for the node itself. In a mesh network, every node is a repeater. In a traditional infrastructure network, nodes must be configured to pass along traffic to other nodes.

A repeater can use one or more wireless devices. When using a single radio (called a ***one-arm repeater***), overall efficiency is slightly less than half of the available bandwidth, since the radio can either send or receive data, but never both at once. These devices are cheaper, simpler, and have lower power requirements. A repeater with two (or more) radio cards can operate all radios at full capacity, as long as they are each configured to use non-overlapping channels. Of course, repeaters can also supply an Ethernet connection to provide local connectivity.

Repeaters can be purchased as a complete hardware solution, or easily assembled by connecting two or more wireless nodes together with Ethernet cable. When planning to use a repeater built with 802.11 technology, remember that nodes must be configured for master, managed, or ad-hoc mode. Typically, both radios in a repeater are configured for master mode, to allow multiple clients to connect to either side of the repeater. But depending on your network layout, one or more devices may need to use ad-hoc or even client mode.

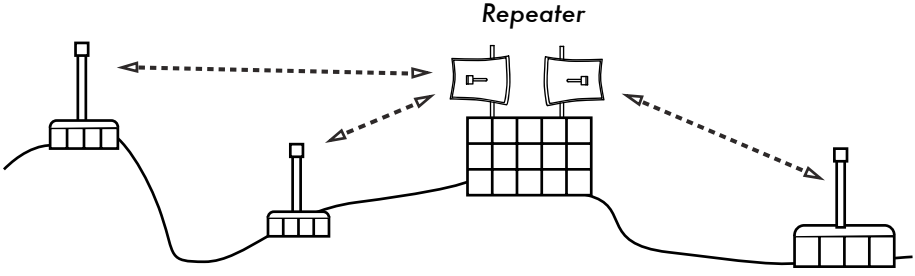


Figure 3.22: The repeater forwards packets over the air between nodes that have no direct line of sight.

Typically, repeaters are used to overcome obstacles in the path of a long distance link. For example, there may be buildings in your path, but those buildings contain people. Arrangements can often be worked out with building owners to provide bandwidth in exchange for roof rights and electricity. If the building owner isn't interested, tenants on high floors may be able to be persuaded to install equipment in a window.

If you can't go over or through an obstacle, you can often go around it. Rather than using a direct link, try a multi-hop approach to avoid the obstacle.

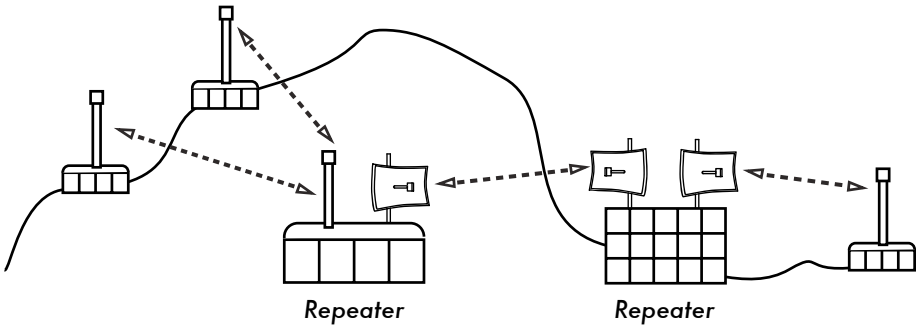


Figure 3.23: No power was available at the top of the hill, but it was circumvented by using multiple repeater sites around the base.

Finally, you may need to consider going backwards in order to go forwards. If there is a high site available in a different direction, and that site can see beyond the obstacle, a stable link can be made via an indirect route.

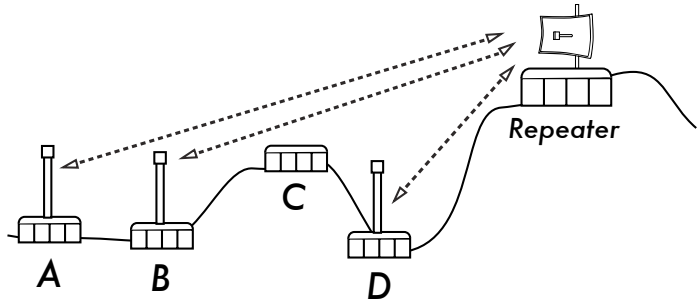


Figure 3.24: Site D could not make a clean link to site A or B, since site C is in the way and is not hosting a node. By installing a high repeater, nodes A, B, and D can communicate with each other. Note that traffic from node D actually travels further away from the rest of the network before the repeater forwards it along.

Repeaters in networks remind me of the "six degrees of separation" principle. This idea says that no matter who you are looking for, you need only contact five intermediaries before finding the person. Repeaters in high places can "see" a great deal of intermediaries, and as long as your node is in range of the repeater, you can communicate with any node the repeater can reach.

Traffic optimization

Bandwidth is measured as the amount of bits transmitted over a time interval. This means that over time, bandwidth available on any link approaches infinity. Unfortunately, for any given period of time, the bandwidth provided by any given network connection is not infinite. You can always download (or upload) as much traffic as you like; you need only wait long enough. Of course, human users are not as patient as computers, and are not willing to

wait an infinite amount of time for their information to traverse the network. For this reason, bandwidth must be managed and prioritized much like any other limited resource.

You will significantly improve response time and maximize available throughput by eliminating unwanted and redundant traffic from your network. This section describes a few common techniques for making sure that your network carries only the traffic that must traverse it. For a more thorough discussion of the complex subject of bandwidth optimization, see the free book *How to Accelerate Your Internet* (*http://bwmo.net/*).

Web caching

A web proxy server is a server on the local network that keeps copies of recently retrieved or often used web pages, or parts of pages. When the next person retrieves these pages, they are served from the local proxy server instead of from the Internet. This results in significantly faster web access in most cases, while reducing overall Internet bandwidth usage. When a proxy server is implemented, the administrator should also be aware that some pages are not cacheable-- for example, pages that are the output of server-side scripts, or other dynamically generated content.

The apparent loading of web pages is also affected. With a slow Internet link, a typical page begins to load slowly, first showing some text and then displaying the graphics one by one. In a network with a proxy server, there could be a delay when nothing seems to happen, and then the page will load almost at once. This happens because the information is sent to the computer so quickly that it spends a perceptible amount of time rendering the page. The overall time it takes to load the whole page might take only ten seconds (whereas without a proxy server, it may take 30 seconds to load the page gradually). But unless this is explained to some impatient users, they may say the proxy server has made things slower. It is usually the task of the network administrator to deal with user perception issues like these.

Proxy server products

There are a number of web proxy servers available. These are the most commonly used software packages:

- **Squid**. Open source Squid is the de facto standard at universities. It is free, reliable, easy to use and can be enhanced (for example, adding content filtering and advertisement blocking). Squid produces logs that can be analyzed using software such as Awstats, or Webalizer, both of which are open source and produce good graphical reports. In most cases, it is easier to install as part of the distribution than to download it from

http://www.squid-cache.org/ (most Linux distributions such as Debian, as well as other versions of Unix such as NetBSD and FreeBSD come with Squid). A good Squid configuration guide can be found on the Squid Users Guide Wiki at *http://www.deckle.co.za/squid-users-guide/*.

- **Microsoft Proxy server 2.0.** Not available for new installations because it has been superseded by Microsoft ISA server and is no longer supported. It is nonetheless used by some institutions, although it should perhaps not be considered for new installations.
- **Microsoft ISA server.** ISA server is a very good proxy server program, that is arguably too expensive for what it does. However, with academic discounts it may be affordable to some institutions. It produces its own graphical reports, but its log files can also be analyzed with popular analyzer software such as Sawmill (*http://www.sawmill.net/*). Administrators at a site with MS ISA Server should spend sufficient time getting the configuration right; otherwise MS ISA Server can itself be a considerable bandwidth user. For example, a default installation can easily consume more bandwidth than the site has used before, because popular pages with short expiry dates (such as news sites) are continually being refreshed. Therefore it is important to get the pre-fetching settings right, and to configure pre-fetching to take place mainly overnight. ISA Server can also be tied to content filtering products such as WebSense. For more information, see: *http://www.microsoft.com/isaserver/* and *http://www.isaserver.org/* .

Preventing users from bypassing the proxy server

While circumventing Internet censorship and restrictive information access policy may be a laudable political effort, proxies and firewalls are necessary tools in areas with extremely limited bandwidth. Without them, the stability and usability of the network are threatened by legitimate users themselves. Techniques for bypassing a proxy server can be found at *http://www.antiproxy.com/* . This site is useful for administrators to see how their network measures up against these techniques.

To enforce use of the caching proxy, you might consider simply setting up a network access policy and trusting your users. In the layout below, the administrator has to trust that his users will not bypass the proxy server.

In this case the administrator typically uses one of the following techniques:

- **Not giving out the default gateway address through DCHP.** This may work for a while, but some network-savvy users who want to bypass the proxy might find or guess the default gateway address. Once that happens, word tends to spread about how to bypass the proxy.

- **Using domain or group policies.** This is very useful for configuring the correct proxy server settings for Internet Explorer on all computers in the domain, but is not very useful for preventing the proxy from being bypassed, because it depends on a user logging on to the NT domain. A user with a Windows 95/98/ME computer can cancel his log-on and then bypass the proxy, and someone who knows a local user password on his Windows NT/2000/XP computer can log on locally and do the same.
- **Begging and fighting with users.** This approach, while common, is never an optimal situation for a network administrator.

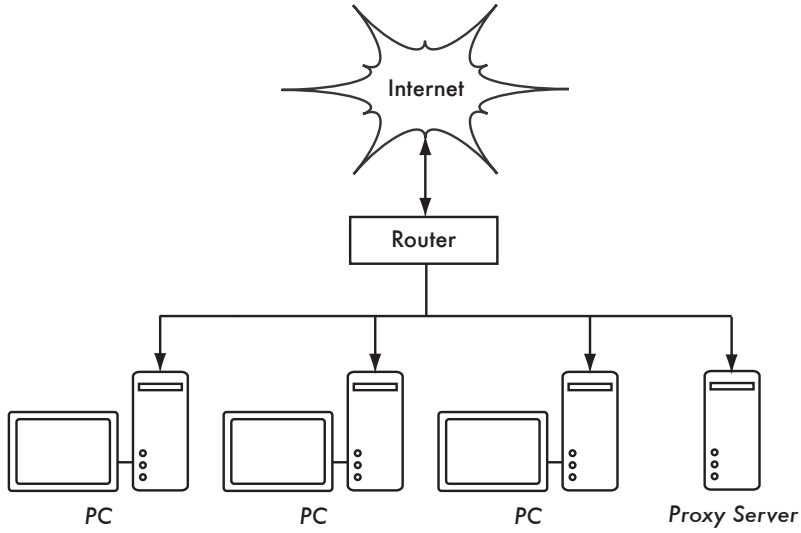


Figure 3.25: This network relies on trusted users to properly configure their PCs to use the proxy server.

The only way to ensure that proxies cannot be bypassed is by using the correct network layout, by using one of the three techniques described below.

Firewall

A more reliable way to ensure that PCs don't bypass the proxy can be implemented using the firewall. The firewall can be configured to allow only the proxy server to make HTTP requests to the Internet. All other PCs are blocked, as shown in **Figure 3.26**.

Relying on a firewall may or may not be sufficient, depending on how the firewall is configured. If it only blocks access from the campus LAN to port 80 on web servers, there will be ways for clever users to find ways around it. Additionally, they will be able to use other bandwidth hungry protocols such as BitTorrent or Kazaa.

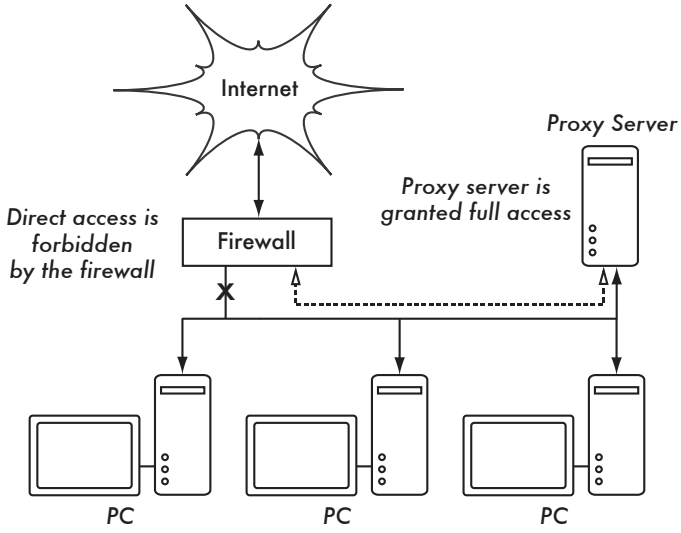


Figure 3.26: The firewall prevents PCs from accessing the Internet directly, but allows access via the proxy server.

Two network cards

Perhaps the most reliable method is to install two network cards in the proxy server and connect the campus network to the Internet as shown below. In this way, the network layout makes it physically impossible to reach the Internet without going through the proxy server.

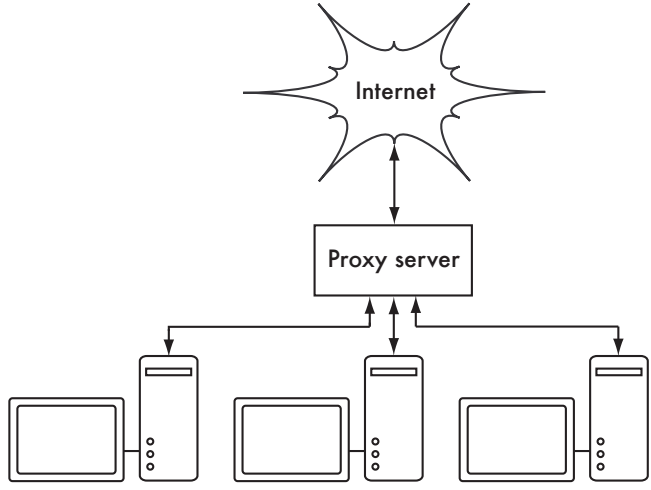


Figure 3.27: The only route to the Internet is through the proxy.

The proxy server in this diagram should not have IP forwarding enabled, unless the administrators knows exactly what they want to let through.

One big advantage to this design is that a technique known as ***transparent proxying*** can be used. Using a transparent proxy means that users' web requests are automatically forwarded to the proxy server, without any need to manually configure web browsers to use it. This effectively forces all web traffic to be cached, eliminates many chances for user error, and will even work with devices that do not support use of a manual proxy. For more details about configuring a transparent proxy with Squid, see:

- *http://www.squid-cache.org/Doc/FAQ/FAQ-17.html*
- *http://tldp.org/HOWTO/TransparentProxy.html*

Policy-based routing

One way to prevent bypassing of the proxy using Cisco equipment is with policy routing. The Cisco router transparently directs web requests to the proxy server. This technique is used at Makerere University. The advantage of this method is that, if the proxy server is down, the policy routes can be temporarily removed, allowing clients to connect directly to the Internet.

Mirroring a website

With permission of the owner or web master of a site, the whole site can be mirrored to a local server overnight, if it is not too large. This is something that might be considered for important websites that are of particular interest to the organization or that are very popular with web users. This may have some use, but it has some potential pitfalls. For example, if the site that is mirrored contains CGI scripts or other dynamic content that require interactive input from the user, this would cause problems. An example is a website that requires people to register online for a conference. If someone registers online on a mirrored server (and the mirrored script works), the organizers of the site will not have the information that the person registered.

Because mirroring a site may infringe copyright, this technique should only be used with permission of the site concerned. If the site runs ***rsync***, the site could be mirrored using rsync. This is likely the fastest and most efficient way to keep site contents synchronized. If the remote web server is not running rsync, the recommended software to use is a program called ***wget***. It is part of most versions of Unix/Linux. A Windows version can be found at *http://xoomer.virgilio.it/hherold/*, or in the free Cygwin Unix tools package (*http://www.cygwin.com/*).

A script can be set up to run every night on a local web server and do the following:

- Change directory to the web server document root: for example, **`/var/www/`** on Unix, or **`C:\Inetpub\wwwroot`** on Windows.
- Mirror the website using the command:

```
wget --cache=off -m http://www.python.org
```

The mirrored website will be in a directory **`www.python.org`**. The web server should now be configured to serve the contents of that directory as a name-based virtual host. Set up the local DNS server to fake an entry for this site. For this to work, client PCs should be configured to use the local DNS server(s) as the primary DNS. (This is advisable in any case, because a local caching DNS server speeds up web response times).

Pre-populate the cache using wget

Instead of setting up a mirrored website as described in the previous section, a better approach is to populate the proxy cache using an automated process. This method has been described by J. J. Eksteen and J. P. L. Cloete of the CSIR in Pretoria, South Africa, in a paper entitled **Enhancing International World Wide Web Access in Mozambique Through the Use of Mirroring and Caching Proxies**. In this paper (available at *http://www.isoc.org/inet97/ans97/cloet.htm*) they describe how the process works:

> "*An automatic process retrieves the site's home page and a specified number of extra pages (by recursively following HTML links on the retrieved pages) through the use of a proxy. Instead of writing the retrieved pages onto the local disk, the mirror process discards the retrieved pages. This is done in order to conserve system resources as well as to avoid possible copyright conflicts. By using the proxy as intermediary, the retrieved pages are guaranteed to be in the cache of the proxy as if a client accessed that page. When a client accesses the retrieved page, it is served from the cache and not over the congested international link. This process can be run in off-peak times in order to maximize bandwidth utilization and not to compete with other access activities.*"

The following command (scheduled to run at night once every day or week) is all that is needed (repeated for every site that needs pre-populating).

```
wget --proxy-on --cache=off --delete after -m http://www.python.org
```

These options enable the following:

- **-m**: Mirrors the entire site. wget starts at *www.python.org* and follows all hyperlinks, so it downloads all subpages.
- **--proxy-on**: Ensures that wget makes use of the proxy server. This might not be needed in set-ups where a transparent proxy is employed.
- **--cache=off**: Ensures that fresh content is retrieved from the Internet, and not from the local proxy server.
- **--delete after**: Deletes the mirrored copy. The mirrored content remains in the proxy cache if there is sufficient disk space, and the proxy server caching parameters are set up correctly.

In addition, wget has many other options; for example, to supply a password for websites that require them. When using this tool, Squid should be configured with sufficient disk space to contain all the pre-populated sites and more (for normal Squid usage involving pages other than the pre-populated ones). Fortunately, disk space is becoming ever cheaper and disk sizes are far larger than ever before. However, this technique can only be used with a few selected sites. These sites should not be too big for the process to finish before the working day starts, and an eye should be kept on disk space.

Cache hierarchies

When an organization has more than one proxy server, the proxies can share cached information among them. For example, if a web page exists in server A's cache, but not in the cache of server B, a user connected via server B might get the cached object from server A via server B. ***Inter-Cache Protocol*** (***ICP***) and ***Cache Array Routing Protocol*** (***CARP***) can share cache information. CARP is considered the better protocol. Squid supports both protocols, and MS ISA Server supports CARP. For more information, see *http://squid-docs.sourceforge.net/latest/html/c2075.html*. This sharing of cached information reduces bandwidth usage in organizations where more than one proxy is used.

Proxy specifications

On a university campus network, there should be more than one proxy server, both for performance and also for redundancy reasons. With today's cheaper and larger disks, powerful proxy servers can be built, with 50 GB or more disk space allocated to the cache. Disk performance is important, therefore the fastest SCSI disks would perform best (although an IDE based cache is better than none at all). RAID or mirroring is not recommended.

It is also recommended that a separate disk be dedicated to the cache. For example, one disk could be for the cache, and a second for the operating system and cache logging. Squid is designed to use as much RAM as it can get, because when data is retrieved from RAM it is much faster than when it

comes from the hard disk. For a campus network, RAM memory should be 1GB or more:

- Apart from the memory required for the operating system and other applications, Squid requires 10 MB of RAM for every 1 GB of disk cache. Therefore, if there is 50 GB of disk space allocated to caching, Squid will require 500 MB extra memory.
- The machine would also require 128 MB for Linux and 128 MB for Xwindows.
- Another 256 MB should be added for other applications and in order that everything can run easily. Nothing increases a machine's performance as much as installing a large amount of memory, because this reduces the need to use the hard disk. Memory is thousands of times faster than a hard disk. Modern operating systems keep frequently accessed data in memory if there is enough RAM available. But they use the page file as an extra memory area when they don't have enough RAM.

DNS caching and optimization

Caching-only DNS servers are not authoritative for any domains, but rather just cache results from queries asked of them by clients. Just like a proxy server that caches popular web pages for a certain time, DNS addresses are cached until their ***time to live*** (***TTL***) expires. This will reduce the amount of DNS traffic on your Internet connection, as the DNS cache may be able to satisfy many of the queries locally. Of course, client computers must be configured to use the caching-only name server as their DNS server. When all clients use this server as their primary DNS server, it will quickly populate a cache of IP addresses to names, so that previously requested names can quickly be resolved. DNS servers that are authoritative for a domain also act as cache name-address mappings of hosts resolved by them.

Bind (named)

Bind is the de facto standard program used for name service on the Internet. When Bind is installed and running, it will act as a caching server (no further configuration is necessary). Bind can be installed from a package such as a Debian package or an RPM. Installing from a package is usually the easiest method. In Debian, type

```
apt-get install bind9
```

In addition to running a cache, Bind can also host authoritative zones, act as a slave to authoritative zones, implement split horizon, and just about everything else that is possible with DNS.

dnsmasq

One alternative caching DNS server is ***dnsmasq***. It is available for BSD and most Linux distributions, or from *http://www.thekelleys.org.uk/dnsmasq/*. The big advantage of dnsmasq is flexibility: it easily acts as both a caching DNS proxy and an authoritative source for hosts and domains, without complicated zone file configuration. Updates can be made to zone data without even restarting the service. It can also serve as a DHCP server, and will integrate DNS service with DHCP host requests. It is very lightweight, stable, and extremely flexible. Bind is likely a better choice for very large networks (more than a couple of hundred nodes), but the simplicity and flexibility of dnsmasq makes it attractive for small to medium sized networks.

Windows NT

To install the DNS service on Windows NT4: select Control Panel ➔ Network ➔ Services ➔ Add ➔ Microsoft DNS server. Insert the Windows NT4 CD when prompted. Configuring a caching-only server in NT is described in Knowledge Base article 167234. From the article:

> "*Simply install DNS and run the Domain Name System Manager. Click on DNS in the menu, select New Server, and type in the IP address of your computer where you have installed DNS. You now have a caching-only DNS server.*"

Windows 2000

Install DNS service: Start ➔ Settings ➔ Control Panel ➔ Add/Remove Software. In Add/Remove Windows Components, select Components ➔ Networking Services ➔ Details ➔ Domain Name System (DNS). Then start the DNS MMC (Start ➔ Programs ➔ Administrative Tools ➔ DNS) From the Action menu select "Connect To Computer..." In the Select Target Computer window, enable "The following computer:" and enter the name of a DNS server you want to cache. If there is a . [dot] in the DNS manager (this appears by default), this means that the DNS server thinks it is the root DNS server of the Internet. It is certainly not. Delete the . [dot] for anything to work.

Split DNS and a mirrored server

The aim of split DNS (also known as ***split horizon***) is to present a different view of your domain to the inside and outside worlds. There is more than one way to do split DNS; but for security reasons, it's recommended that you have two separate internal and external content DNS servers (each with different databases).

Split DNS can enable clients from a campus network to resolve IP addresses for the campus domain to local RFC1918 IP addresses, while the rest of the

Internet resolves the same names to different IP addresses. This is achieved by having two zones on two different DNS servers for the same domain.

One of the zones is used by internal network clients and the other by users on the Internet. For example, in the network below the user on the Makerere campus gets *http://www.makerere.ac.ug/* resolved to 172.16.16.21, whereas a user elsewhere on the Internet gets it resolved to 195.171.16.13.

The DNS server on the campus in the above diagram has a zone file for *makerere.ac.ug* and is configured as if it is authoritative for that domain. In addition, it serves as the DNS caching server for the Makerere campus, and all computers on the campus are configured to use it as their DNS server.

The DNS records for the campus DNS server would look like this:

```
makerere.ac.ug
www CNAME      webserver.makerere.ac.ug
ftp CNAME      ftpserver.makerere.ac.ug
mail CNAME     exchange.makerere.ac.ug
mailserver     A       172.16.16.21
webserver      A       172.16.16.21
ftpserver      A       172.16.16.21
```

But there is another DNS server on the Internet that is actually authoritative for the *makerere.ac.ug* domain. The DNS records for this external zone would look like this:

```
makerere.ac.ug
www A 195.171.16.13
ftp A 195.171.16.13
mail    A 16.132.33.21
        MX mail.makerere.ac.ug
```

Split DNS is not dependent on using RFC 1918 addresses. An African ISP might, for example, host websites on behalf of a university but also mirror those same websites in Europe. Whenever clients of that ISP access the website, it gets the IP address at the African ISP, and so the traffic stays in the same country. When visitors from other countries access that website, they get the IP address of the mirrored web server in Europe. In this way, international visitors do not congest the ISP's VSAT connection when visiting the university's website. This is becoming an attractive solution, as web hosting close to the Internet backbone has become very cheap.

Internet link optimization

As mentioned earlier, network throughput of up to 22 Mbps can be achieved by using standard, unlicensed 802.11g wireless gear. This amount of bandwidth will likely be at least an order of magnitude higher than that provided by

your Internet link, and should be able to comfortably support many simultaneous Internet users.

But if your primary Internet connection is through a VSAT link, you will encounter some performance issues if you rely on default TCP/IP parameters. By optimizing your VSAT link, you can significantly improve response times when accessing Internet hosts.

TCP/IP factors over a satellite connection

A VSAT is often referred to as a ***long fat pipe network***. This term refers to factors that affect TCP/IP performance on any network that has relatively large bandwidth, but high latency. Most Internet connections in Africa and other parts of the developing world are via VSAT. Therefore, even if a university gets its connection via an ISP, this section might apply if the ISP's connection is via VSAT. The high latency in satellite networks is due to the long distance to the satellite and the constant speed of light. This distance adds about 520 ms to a packet's round-trip time (RTT), compared to a typical RTT between Europe and the USA of about 140 ms.

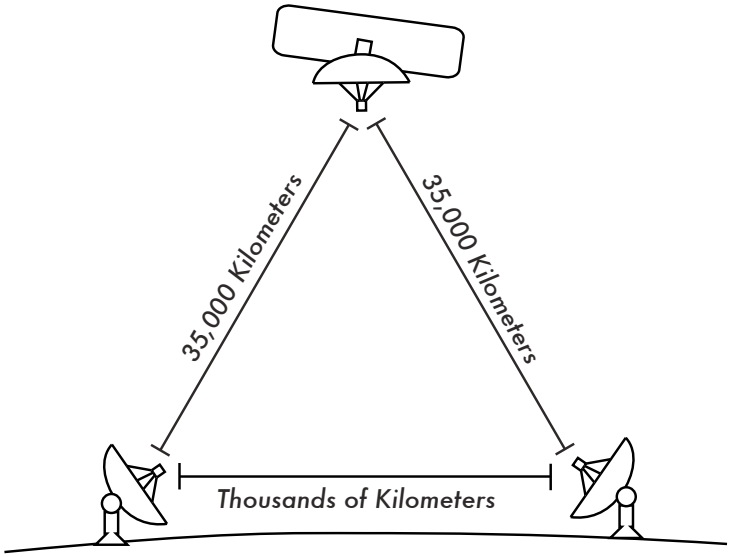


Figure 3.28: Due to the speed of light and long distances involved, a single ping packet can take more than 520 ms to be acknowledged over a VSAT link.

The factors that most significantly impact TCP/IP performance are **long RTT**, **large bandwidth delay product**, and **transmission errors**.

Generally speaking, operating systems that support modern TCP/IP implementations should be used in a satellite network. These implementations support the RFC 1323 extensions:

- The ***window scale*** option for supporting large TCP window sizes (larger than 64KB).
- ***Selective acknowledgment*** (***SACK***) to enable faster recovery from transmission errors.
- Timestamps for calculating appropriate RTT and retransmission timeout values for the link in use.

Long round-trip time (RTT)

Satellite links have an average RTT of around 520ms to the first hop. TCP uses the slow-start mechanism at the start of a connection to find the appropriate TCP/IP parameters for that connection. Time spent in the slow-start stage is proportional to the RTT, and for a satellite link it means that TCP stays in slow-start mode for a longer time than would otherwise be the case. This drastically decreases the throughput of short-duration TCP connections. This is can be seen in the way that a small website might take surprisingly long to load, but when a large file is transferred acceptable data rates are achieved after a while.

Furthermore, when packets are lost, TCP enters the congestion-control phase, and owing to the higher RTT, remains in this phase for a longer time, thus reducing the throughput of both short- and long-duration TCP connections.

Large bandwidth-delay product

The amount of data in transit on a link at any point of time is the product of bandwidth and the RTT. Because of the high latency of the satellite link, the bandwidth-delay product is large. TCP/IP allows the remote host to send a certain amount of data in advance without acknowledgment. An acknowledgment is usually required for all incoming data on a TCP/IP connection. However, the remote host is always allowed to send a certain amount of data without acknowledgment, which is important to achieve a good transfer rate on large bandwidth-delay product connections. This amount of data is called the ***TCP window size***. The window size is usually 64KB in modern TCP/IP implementations.

On satellite networks, the value of the bandwidth-delay product is important. To utilize the link fully, the window size of the connection should be equal to the bandwidth-delay product. If the largest window size allowed is 64KB, the maximum theoretical throughput achievable via satellite is (window size) / RTT, or 64KB / 520 ms. This gives a maximum data rate of 123 KB/s, which is 984 kbps, regardless of the fact that the capacity of the link may be much greater.

Each TCP segment header contains a field called ***advertised window***, which specifies how many additional bytes of data the receiver is prepared to accept. The advertised window is the receiver's current available buffer size.

The sender is not allowed to send more bytes than the advertised window. To maximize performance, the sender should set its send buffer size and the receiver should set its receive buffer size to no less than the bandwidth-delay product. This buffer size has a maximum value of 64KB in most modern TCP/IP implementations.

To overcome the problem of TCP/IP stacks from operating systems that don't increase the window size beyond 64KB, a technique known as ***TCP acknowledgment spoofing*** can be used (see Performance Enhancing Proxy, below).

Transmission errors

In older TCP/IP implementations, packet loss is always considered to have been caused by congestion (as opposed to link errors). When this happens, TCP performs congestion avoidance, requiring three duplicate ACKs or slow start in the case of a timeout. Because of the long RTT value, once this congestion-control phase is started, TCP/IP on satellite links will take a longer time to return to the previous throughput level. Therefore errors on a satellite link have a more serious effect on the performance of TCP than over low latency links. To overcome this limitation, mechanisms such as ***Selective Acknowledgment*** (***SACK***) have been developed. SACK specifies exactly those packets that have been received, allowing the sender to retransmit only those segments that are missing because of link errors.

The Microsoft Windows 2000 TCP/IP Implementation Details White Paper states

> "*Windows 2000 introduces support for an important performance feature known as Selective Acknowledgment (SACK). SACK is especially important for connections using large TCP window sizes.*"

SACK has been a standard feature in Linux and BSD kernels for quite some time. Be sure that your Internet router and your ISP's remote side both support SACK.

Implications for universities

If a site has a 512 kbps connection to the Internet, the default TCP/IP settings are likely sufficient, because a 64 KB window size can fill up to 984 kbps. But if the university has more than 984 kbps, it might in some cases not get the full bandwidth of the available link due to the "long fat pipe network" factors discussed above. What these factors really imply is that they prevent a single machine from filling the entire bandwidth. This is not a bad thing during the day, because many people are using the bandwidth. But if, for example, there are large scheduled downloads at night, the administrator might want those downloads to make use of the full bandwidth, and the "long fat pipe network" factors might be an obstacle. This may also become critical

if a significant amount of your network traffic routes through a single tunnel or VPN connection to the other end of the VSAT link.

Administrators might consider taking steps to ensure that the full bandwidth can be achieved by tuning their TCP/IP settings. If a university has implemented a network where all traffic has to go through the proxy (enforced by network layout), then the only machines that make connections to the Internet will be the proxy and mail servers.

For more information, see *http://www.psc.edu/networking/perf_tune.html* .

Performance-enhancing proxy (PEP)

The idea of a Performance-enhancing proxy is described in RFC 3135 (see *http://www.ietf.org/rfc/rfc3135*), and would be a proxy server with a large disk cache that has RFC 1323 extensions, among other features. A laptop has a TCP session with the PEP at the ISP. That PEP, and the one at the satellite provider, communicate using a different TCP session or even their own proprietary protocol. The PEP at the satellite provider gets the files from the web server. In this way, the TCP session is split, and thus the link characteristics that affect protocol performance (long fat pipe factors) are overcome (by TCP acknowledgment spoofing, for example). Additionally, the PEP makes use of proxying and pre-fetching to accelerate web access further.

Such a system can be built from scratch using Squid, for example, or purchased "off the shelf" from a number of vendors.

More information

While bandwidth optimization is a complex and often difficult subject, the techniques in this chapter should help reduce obvious sources of wasted bandwidth. To make the best possible use of available bandwidth, you will need to define a good access policy, set up comprehensive monitoring and analysis tools, and implement a network architecture that enforces desired usage limits.

For more information about bandwidth optimization, see the free book *How to Accelerate Your Internet* (*http://bwmo.net/*).