E-4

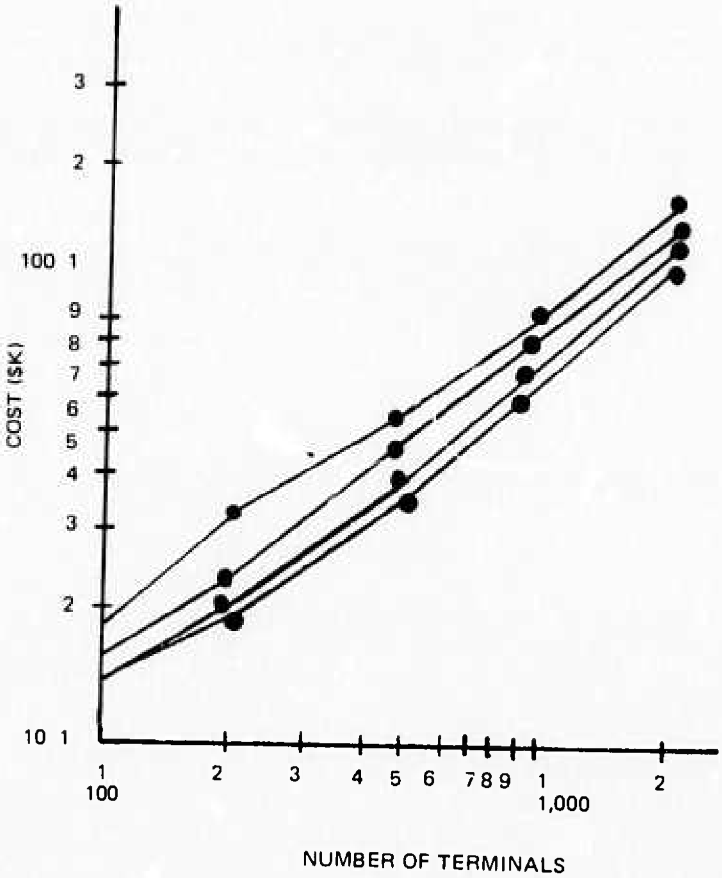


Figure 7.3: Preliminary Estimates of Terminal Connection Costs

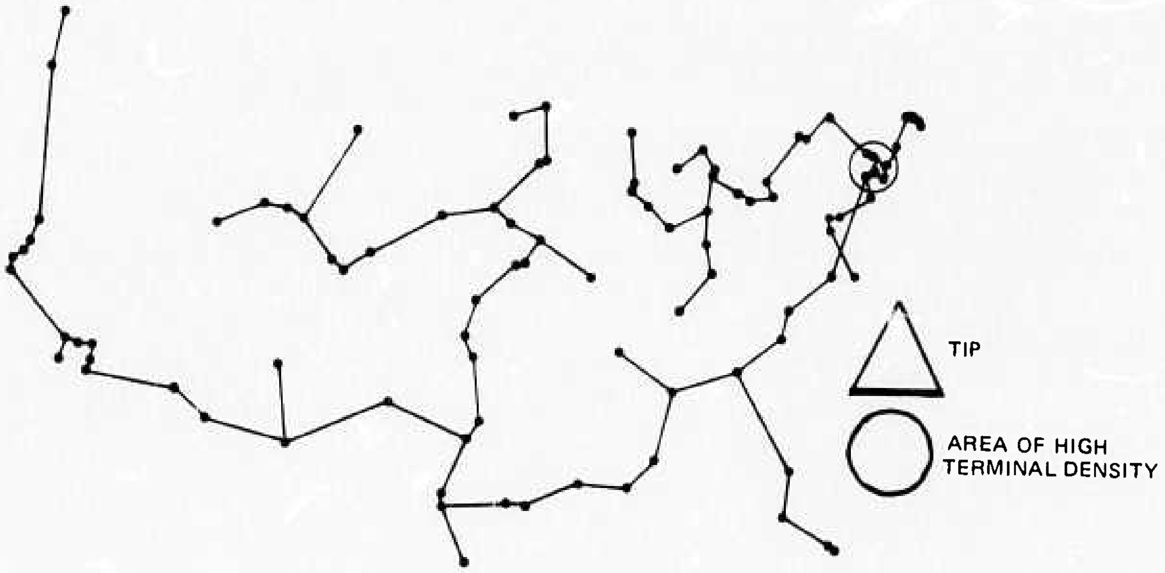


Figure 7.4: Network Design for 100 Nodes, 10bps Traffic, TIP in N.Y.C.

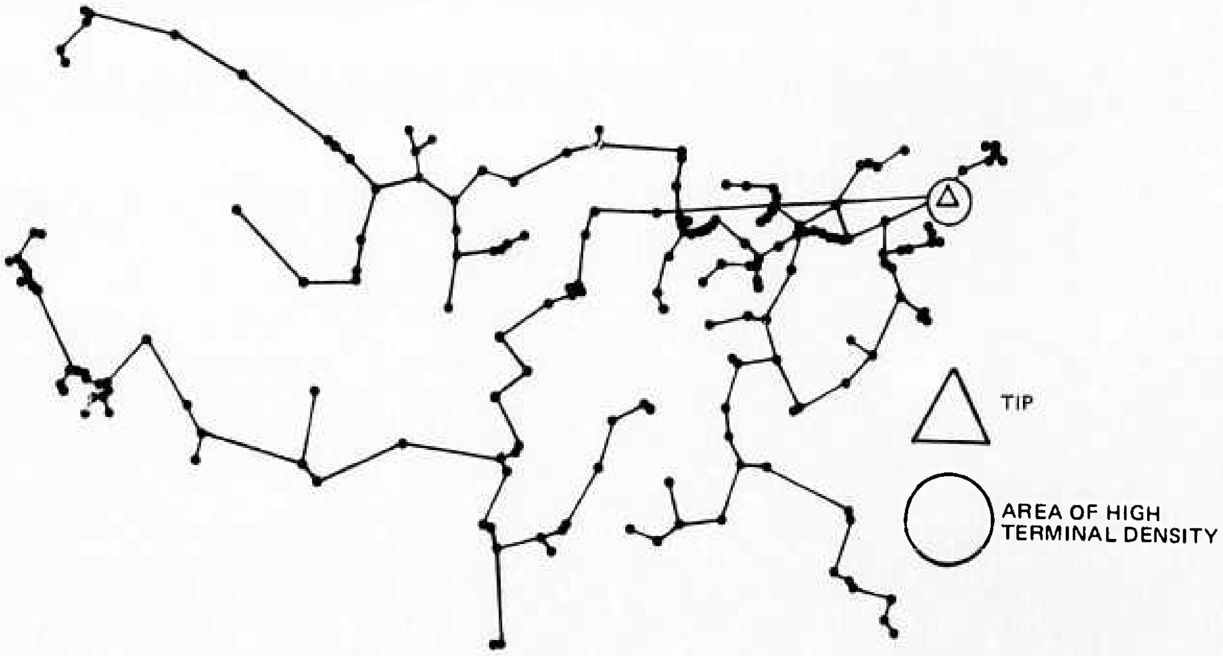


Figure 7.5: Network Design for 200 Nodes, 10bps Traffic, TIP in N.Y.C.

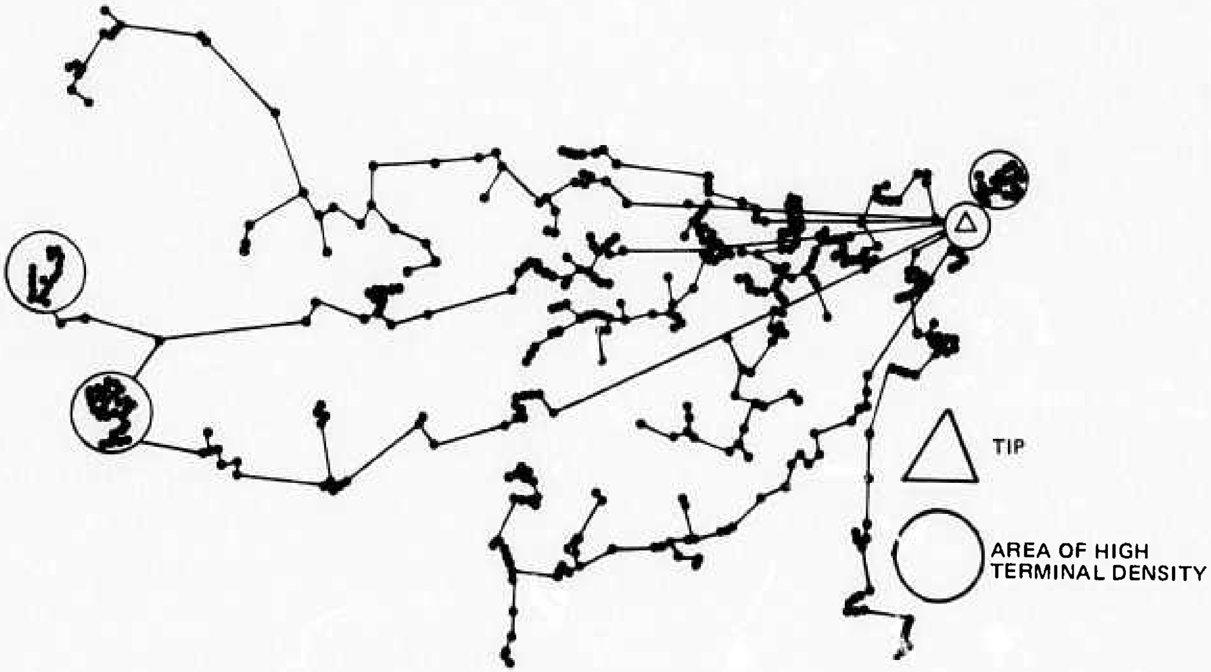


Figure 7.6: Network Design for 500 Nodes, 10bps Traffic, TIP in N.Y.C.

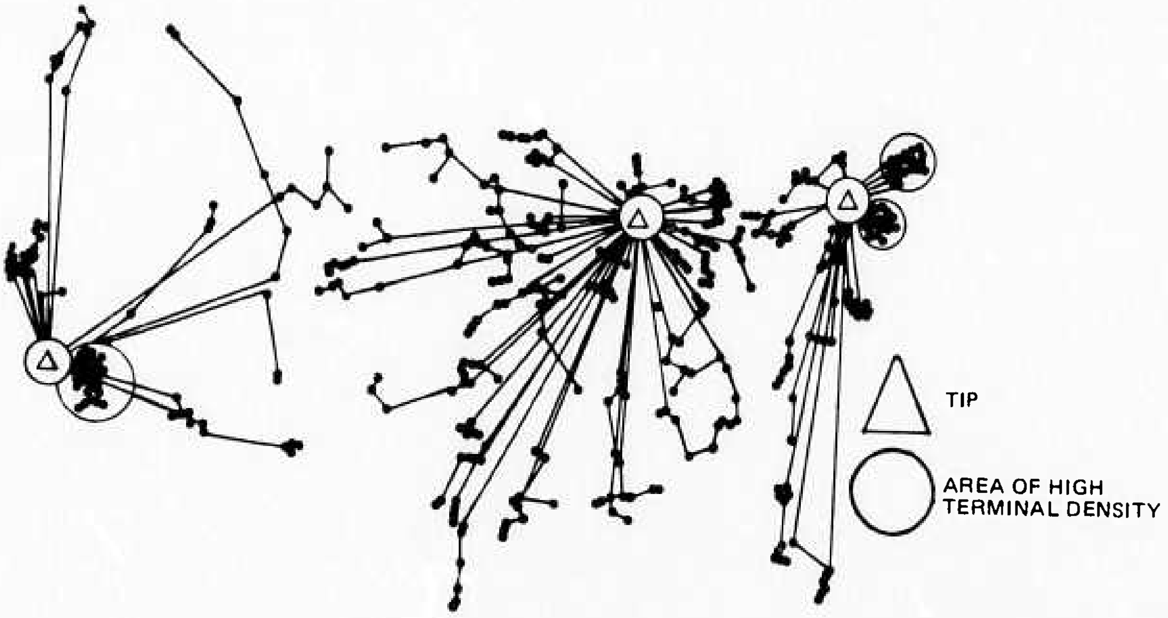


Figure 7.7: Network Design for 500 Nodes, 100bps Traffic, TIP in N.Y.C., Chicago, L.A.

Chapter 8
LOCAL ACCESS—A RING DESIGN EXAMPLE

For local transmission of signals from a nationwide interconnecting network, the user's technical problems are complex because many of the techniques are in the experimental stage. The problem is not just one of configuring facilities, but actually designing the channel. The classical technique of using multidrop lines with polling concentrators as described in Chapter 7 is available and in many cases the best strategy. But, new techniques such as the use of rings or random access multiplexing offer better prospects in many cases. However, neither of these are standard techniques and hence protocols and hardware are in a developmental stage. Furthermore, new physical links are becoming available. One of the most promising of these is the coaxial cable of cable television (CATV) systems.

To illustrate some of the complexities and surprises awaiting the designer of local systems, we present one example of a ring design. In Chapter 9, we present a detailed consideration of the use of CATV systems for local data transmission. In the remainder of this report, Chapters 10 through 15, we discuss the use of broadcast packet radio techniques for handling the local access problem.

Let us illustrate just one of the problems with a ring network, inflexibility in routing that results because there are no alternate routes. The analysis will show that although the ring may accommodate a large throughput when high traffic points are close on the ring, there is no flexibility in adapting to redistribution of traffic requirements. The example is carried out for a mixture of tape transfers and interactive traffic.

One of the traffic models developed by Hayes and Sherman [31] is used to analyze the ring design. We consider the design of a single slotted ring to which sources of traffic are connected through an interface. The source can represent host computers, terminals, or a combination of these. The interface is assumed to receive packets from the source, store them, and multiplex them onto the ring. A header which addresses the packet to a particular interface on the ring is added to the packet; the packet size on the ring is therefore larger than that on the line. In the reverse direction, the interface removes packets from the ring addressed to it, removes the "ring header," and transmits these packets to the source. It is assumed that an interface can remove a packet from the ring and then feed a new packet into that same slot. In this case, the traffic on the ring seen by the interface is in the location marked by X in Figure 8.1. That is, it includes only the traffic

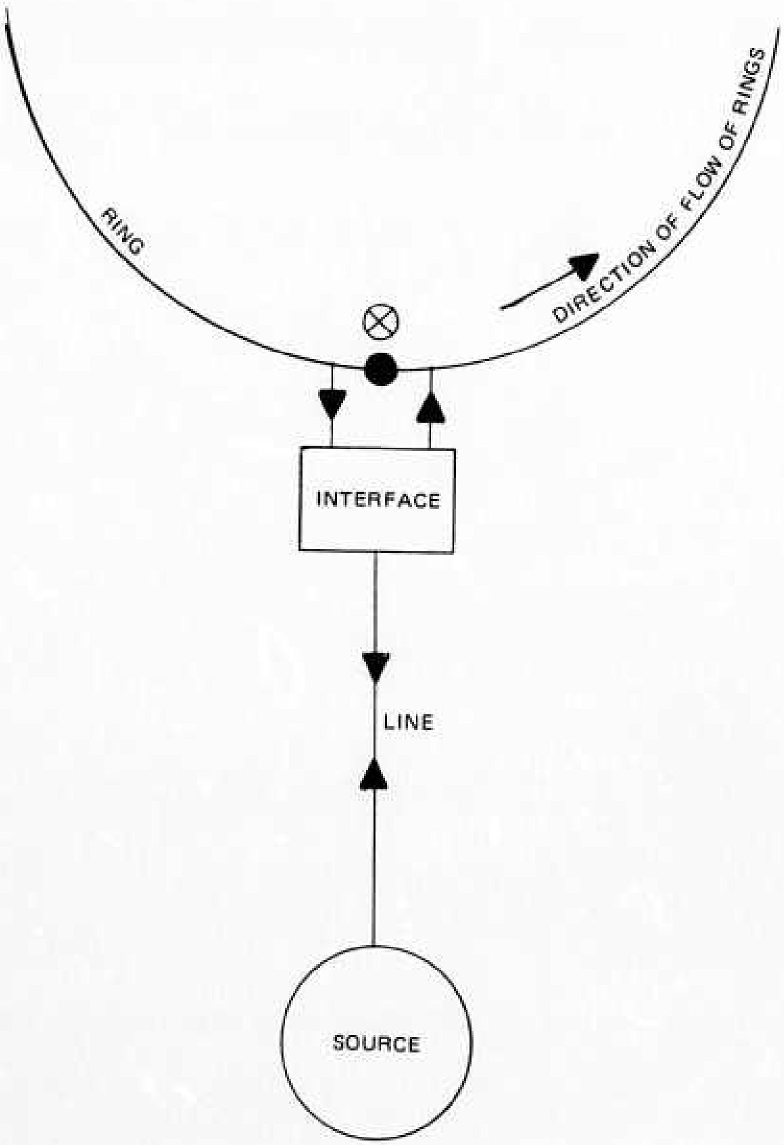


Figure 8.1: Ring-Source Interface

which passes through the interface and not the traffic which is destined for that interface or which originates from it. It is assumed that durations of idle periods on the ring are exponentially distributed.

In the calculations of the buffer content and the delay, we assume that the traffic flow from a source to its interface is at a constant rate, equal to the average rate. The performance of the system is characterized by the buffer size at the interfaces, the delays at these interfaces, and maximum throughput which can be obtained.

For all cases, we used 1.5 Mbps speed on the ring, a packet of 784 bits on the ring, and 768 bits on the lines to the interfaces. Table 8.1 gives the input data to the interfaces, and Table 8.2 gives the distribution matrix P_{ij}, that is the fraction of traffic from interface i destined to interface j. An important parameter is the average number of packets per message. We assume an average of 14 packets/message when all traffic is of an interactive type, 65105 packets/message when all of the traffic is tape transfer, and an average of (1500 ÷ 65105) packets/message for other interfaces depending on the fraction of tape transfers that it included.

For the given data we analyze two designs referred to as System 1 and System 2. System 1 connects the interfaces in order 1 through 4, and the direction of flow on the ring is counter clockwise. Interface pairs with high traffic requirements are relatively close. For System 2, the ring is reconfigured for flow in the following direction: 1,14,8,6,7,9,12,10, 4,13,11,3,2,5,1.

Tables 8.3a and 8.3b show the results for System 1 and 2. Each table shows utilization of the ring seen by an interface, the rate of packets/sec. on the ring seen by an interface,

Table 8.1: Data at Interface

Interface No.	Line Speed Bits/Sec	Rate In Bits/Sec	Rate In Rate/Sec	Average No. of Packets/Msg.	Ratio of Source To Ring Rate
1	230000.	87754.	114.3	14974.	.15655
2	100000.	59554.	77.5	65105.	.06807
3	100000.	61646.	80.3	14.	.07071
4	100000.	41554.	54.1	65105.	.06807
5	100000.	30785.	40.1	26042.	.06807
6	100000.	10154.	13.2	65105.	.06807
7	100000.	17754.	23.1	65105.	.06807
8	230000.	133754.	174.2	29948.	.15655
9	100000.	17015.	22.2	26042.	.06807
10	230000.	69754.	90.8	44922.	.15655
11	100000.	30154.	39.3	14.	.07071
12	100000.	41015.	53.4	26042.	.06807
13	100000.	25354.	33.0	14.	.07071
14	100000.	65554.	85.4	65105.	.06807

Table 8.2: Fraction of Traffic To Interface

Fraction of Traffic to NIP													
1	2	3	4	5	6	7	8	9	10	11	12	13	14
0.000	.091	.227	.066	.063	0.000	0.000	.340	.031	.055	0.000	.072	0.000	.055
.336	0.000	.358	0.000	.013	0.000	0.000	0.000	.026	.081	0.000	.026	.161	0.000
.082	.815	0.000	.004	.016	.004	.004	.004	.029	.004	.004	.029	.004	.004
.116	0.000	.056	0.000	.043	0.000	0.000	0.000	.037	.116	0.000	.037	.597	0.000
.158	.002	.077	.002	0.000	.002	.327	.002	.052	.158	.080	.052	.002	.080
0.000	0.000	.227	0.000	.076	0.000	.394	0.000	.152	0.000	0.000	.152	0.000	0.000
0.000	0.000	.130	0.000	.043	0.000	0.000	0.000	.199	.541	0.000	.087	0.000	0.000
.223	0.000	.017	0.000	.006	0.000	.486	0.000	.012	.022	0.000	.012	0.000	.223
.080	.009	.145	.009	.054	.009	.009	.009	0.000	.268	.009	.241	.009	.150
.069	.049	.067	.275	.080	0.000	.138	0.000	.036	0.000	.112	.105	.069	0.000
.318	0.000	.077	.080	.105	0.000	0.000	0.000	.051	.159	0.000	.131	0.000	.080
.121	.004	.060	.004	.023	.004	.004	.004	.041	.121	.062	0.000	.491	.062
0.000	.379	.091	.189	.030	0.000	0.000	0.000	.061	.189	0.000	.061	0.000	0.000
.073	0.000	.035	.037	.048	0.000	0.000	.381	.023	0.000	.037	.060	.305	0.000

Table 8.3: Comparison of Ring Designs

	Interface Number	Utilization Of Ring Seen By Interface	Pack/Sec On Ring Seen By Interface	Average Number of Packets In Interface Queue	Average Delay Per Packet In Seconds
Table 8.3a System 1	4	.2215	423.04	.44	.00821
	13	.2288	437.60	.28	.00854
Table 8.3b System 2	4	.1811	346.41	40.28	.74447
	13	.1550	296.59	29.95	.90726

the average number of packets waiting to be multiplexed onto the ring, and the average delay per packet. An important point to notice is that in System 2 at interface 4, the average number of packets in the queue is over 40, a severe degradation in performance caused by a redistribution in traffic requirements.

Chapter 9
CATV SYSTEMS FOR LOCAL ACCESS

9.1 Introduction

A wide variety of system configurations such as loop structures and various multiplexing schemes have been proposed for communicating data on future CATV Systems [39].

A circuit switched video system has been developed by Rediffusion International Ltd. in Great Britain [24]. Multipair cables are used with each pair being dedicated to a separate subscriber. He may then select the program of his choice by means of a telephone type dial. The Rediffusion System presents interesting tradeoffs between initial investment, flexibility and reliability. However, since this type of system has not made significant inroads into the U.S. market at present we will not consider it further here.

We first present a very brief introduction to the structure common to most of the 3000 current U.S. CATV Systems [57].

Signals are received at an antenna located for ideal reception and are then relayed from this "head end" to individual subscribers via a distribution system of coaxial cables, broadband repeater amplifiers, and subscriber taps.

A cable television distribution system generally consists of a trunk section and a feeder section. The trunk section contains trunk cable connecting the head end to distribution points, from which the feeder cable emanates. Located along the trunk cable are high-quality repeater amplifiers, which provide gain along the trunk and to the feeders. At the termination of the trunks there are distribution amplifiers. Along the feeder cable there are lower quality amplifiers called extender amplifiers and subscriber taps that provide signals to drop cables leading to home receivers.

With recent broadband amplifiers, the full Sub-UHF spectrum from 5 to 300 MHZ has been used. Partitioned into 6 MHZ channels for television, only a small amount of this spectrum is currently used for TV signals.

FCC regulations now require that new CATV Systems must have two-way capability. Practically speaking, this does not mean that all new systems are two-way systems, but rather that amplifier units are installed with forward amplifier modules in place and with distances between amplifiers constrained so that at some future date reverse amplifier modules can be installed for two-way operation. However, a number of actual fully two-way systems are presently being built and the number is increasing rapidly. Most present two-way systems use the configuration in Figure 9.1a. Filters at each end of the station separate low (L) and high (H) frequencies and direct them to amplifiers. Two possible "two-way" configurations [33] are shown in Figures 9.1b and c.

Of course, two-way CATV Systems are themselves in an experimental stage so that there are still implementation problems in achieving written specifications. Some of the classical electrical and communication bugs are being removed at present—ringing around loops through band separation filters, tuning return AGC's, alignment procedures and construction problems.

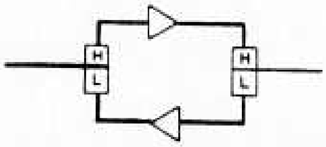


Figure 9.1a: Two-Way CATV Repeater

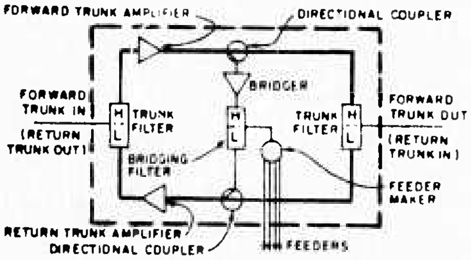


Figure 9.1b: Two-Way CATV Repeater (With Feeders)

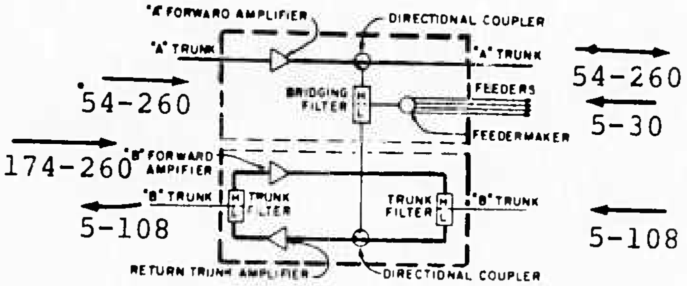


Figure 9.1c: Dual Trunk/Single Feeder Station
(Suburban Boston Configuration)

A number of companies have developed system concepts and subscriber hardware to implement digital home response modes for existing CATV Systems [7, 8]. Among these are Theta-Com, Jerrold, Rediffusion Electronics, CAS Manufacturing Co., Hughes Aircraft, AMECO, Scientific Atlanta, and Cable Information Systems. Several of these companies are running prototype systems in cities throughout the U.S.—El Segundo, California; Dennisport, Mass.; and Orlando, Florida, among them. In addition MITRE of McLean, Va., has installed an experimental system in Reston, Va., which incorporates a "frame grabbing facility" to enable the viewer to store a frame of video data produced by a character generator. Data frames are sent every 1/60th of a second interlaced with standard video frames [56].

In most of these systems an FSK or PSK signal occupying a 4 MHZ bandwidth is used at about a 1 megabit per second rate with different carrier frequencies to and from the central antenna site. In each case, customers are polled at regular intervals to determine access to the channel. Typical proposed uses of these systems are opinion polling, meter reading, shopping, systems diagnostics and alarms. Acceptable response times are in the order of several seconds or in some cases, even minutes [30]. We will investigate data transmission on existing CATV systems with required response times of tenths of seconds and with up to 100,000 interactive users.

To illustrate our points in detail, we will consider a specific design for the Suburban Boston complex. The design will use the "feeder backer" configuration shown in Figure 9.1.c with the frequencies assigned as specificly indicated. The design techniques for the CATV Systems themselves are well known applications of classical communications techniques [16, 21].

9.2 Data Error Rates on CATV Systems

Two-way CATV systems permit input from virtually any location in the network. The result is a large number of noise sources being fed upstream toward a common source. CATV amplifiers have a noise figure of about 10db for a 5 MHz channel. Cascading amplifiers can increase effective system noise figure by 30db or more. Nevertheless, we shall see that system specifications on signal-to-noise ratio for CATV systems are stringent enough so that data can be sent with existing analog repeaters, and no digital repeaters, such that bit rate error probabilities are negligible.

For example, if the worst signal-to-thermal noise ratio is limited to 43db and the worst cross-modulation to signal ratio is limited to –47db, system operators may want to limit data channel carriers to a level of 10 to 20db below TV operating levels in order to minimize additional loading due to the data channel carriers [51]. Accepting these restrictions, in the worst case, we would be limited to 23db signal to thermal noise ratio and -27db cross-modulation to signal ratio. Let us consider both of these sets of restrictions to determine the resulting CATV system performance for random access packet transmission.

We calculate error rates for a FSK system with incoherent detection to determine a lower bound for system performance. The error rates for coherent detection or phase shift keying, of course, would be even lower.

Let S = Signal power
N = Noise power
N_c = Cross-modulation noise power
N_r = Thermal noise power
t = Average synchronization error time
T = Bit width time

Then the sign to noise ratio is:

$$\frac{S}{N} = \frac{Sq}{N_c + N_r} = \frac{q}{(N_{c/s}) + (N_r/S)}$$

where $q = (1 - \frac{2t}{T})^2$

Let P_e be the bit rate error probability.
Let m be the number of keying frequencies in a multiple FSK system.
Then [49]:

$$P_e = \frac{m-1}{m} e^{\frac{-(S/N)}{2(m-1)^2}}$$

We assume that each packet carries its own synchronizing bits and hence there is no need to synchronize every terminal to a master clock. Therefore, temperature, pressure, and humidity variations which have approximately the same effects at all frequencies do not enter into the calculation of t. The group delay variation over a six Megahertz bandwidth is less than .2 μseconds [48]. For a 1 Megabit pulse rate T = 1 μsecond and t = .2 μseconds. Hence, q = .36. We, therefore, have the error probabilities in Table 9.1 for the suburban Boston complex.

For effective signal-to-noise ratios above 20db there is a threshold effect for error probabilities. This is borne out by the negligible error rates. Even for the degraded specifications the error rate is low enough for the most stringent practical data requirements. Furthermore, at a rate of 10^6 pulses/second the FSK signal will occupy the 6MHz bandwidth with negligible intermodulation into TV channels.

Consideration of reflections, intersymbol interference and 60 cycle hum also lead to the conclusion that CATV systems are excellent media for packet data transmission.

The signal levels in a CATV system are controlled via AGC and dual pilot carriers. Ripples are kept to less than 1db over the whole frequency band. In any case, frequency shift

Table 9.1: Error Rates for FSK

System Label	Type of Specification	(N_c/S)	(S/N_r)	m	P_e
A	Undegraded Specs.	-47db	43db	2	$1/2e^{-5,148} \approx 1/2 \times 10^{-2,239}$
B	Undegraded Specs.	-47db	43db	4	$3/4e^{-571} \approx 3/4 \times 10^{-248}$
C	Undegraded Specs.	-47db	43db	8	$7/8e^{-104} \approx 7/8 \times 10^{-45}$
D	Boston Specs. - degraded by 20db	-27db	23db	2	$1/2e^{-51} \approx .5 \times 10^{-22}$

keying is insensitive to small amplitude variations. The effect of group delay error has already been taken into account in the use of q in the formula for error probability. The remaining source of intersymbol interference is the reflection of pulses and the effect of the reflected pulses on the transmitted data. There are three types of disturbances due to reflections. In each case, we shall see that video restrictions are certainly stringent enough to avoid any difficulties for data transmission.

Periodic changes of minute magnitude uniformly distributed along the cable length, the magnitude of changes being essentially equal from period to period, due to the nature of the manufacturing process, cause reflections which add in phase at certain frequencies. The signal strength relationship of the reflected wave to the incident wave is referred to as structural return loss (SRL). Typical values for the magnitude of SRL are better than -26db [43].

Assuming that the reflected signal is always of an opposite sign to the original signal, the signal level is degraded by at most S–a, where a is the amplitude of the reflected signal. The signal-to-noise ratio becomes [53]:

$$\frac{S-a}{N} = \frac{S}{N}\left(1 - \frac{a}{S}\right)$$

In other words $\frac{S}{N}$ is degraded by $\left(1 - \frac{a}{S}\right)$

For a reflected signal of -26db, $\left(1 - \frac{a}{S}\right)$ is .9975—quite acceptable.

A localized change or changes on the cable cause echo phenomena. Low reflection coefficients of active and passive devices and the use of directional couplers at all subscriber taps ensure that the magnitudes of reflected pulses are in the "no ghost range" of Figure 9.2 [45, 40]. These are translated into critical distances for different types of cable in Figure 9.3. Thus, for example, considering the reflection on .412 inch cable at Channel 13 the critical distance is about 250 feet and the ratio of the magnitude of the reflected signal to the magnitude of the original signal is -23db.

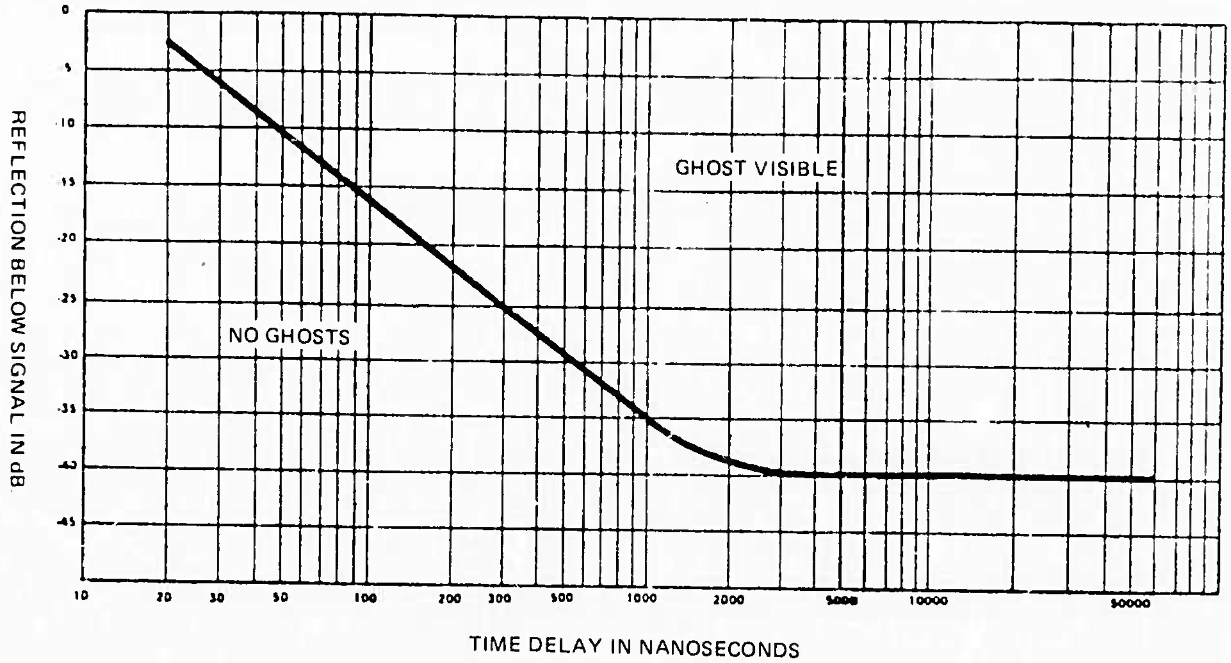


Figure 9.2: Curve Showing Perceptibility of Ghosts

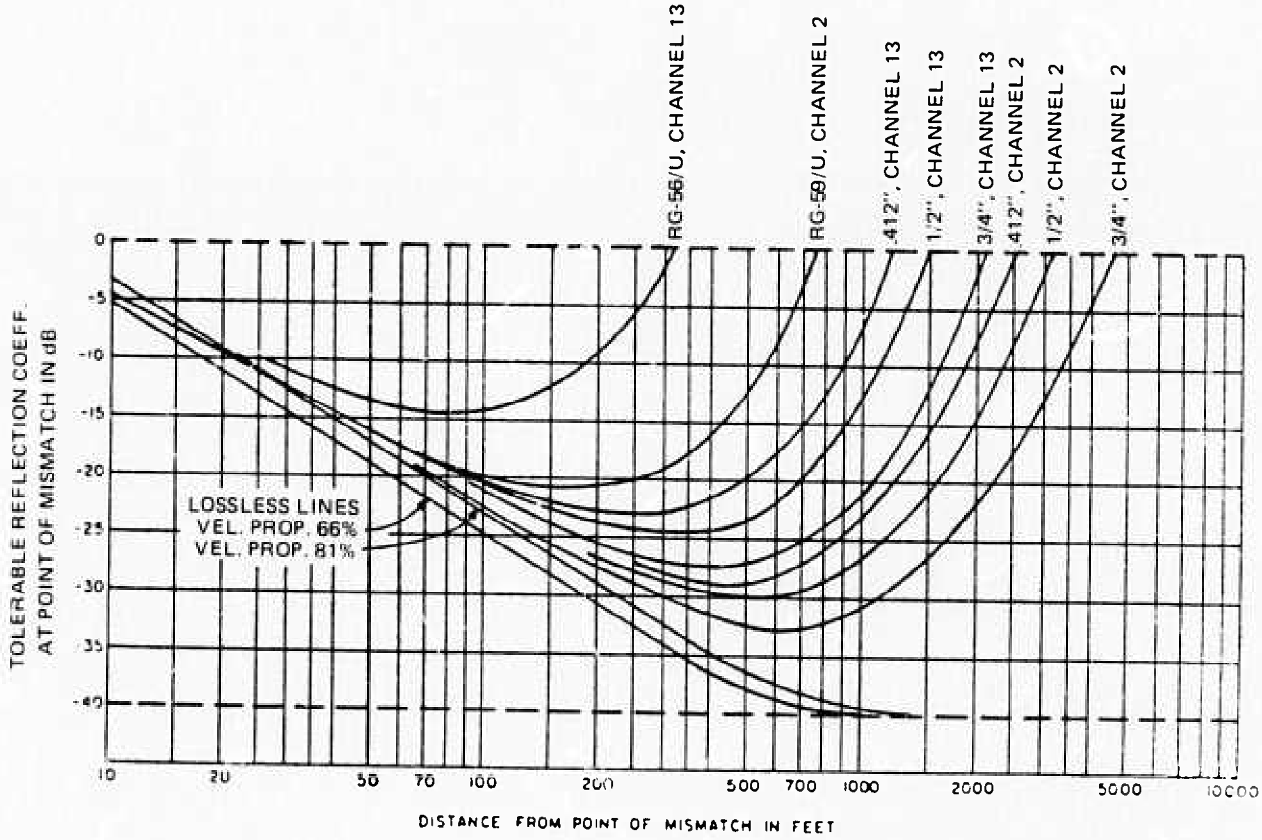


Figure 9.3: Graph for Determination of Critical Cable Lengths

Randomly distributed changes of random magnitude which persist throughout the cable length cause reflections which do not add in phase. These can be taken into account in noise calculations and are usually negligible.

Cable system amplifiers are powered by low voltage 60 Hz power through the co-axial cable. This power may be as high as 60 volts (RMS) and currents may run to 10 ampheres (RMS) with peak currents even higher. There are significant harmonics of the power line frequencies present. Some amplifiers use switching mode power supplies with switching frequencies in the 10-20 KHz range. Hash from these switching regulators also finds its way into the cable. However, both the 60 cycle harmonics and hash limit only the area of very low frequencies which are generally avoided for data transmission.

9.3 Other Performance Criteria for CATV Systems

Data users may find cable system reliability quite poor when compared with the common carrier facilities with which they are familiar. One of the major problems with data transmission on CATV Systems is that there is no redundancy of path cables or amplifiers. There are no government or industry minimal standards for acceptable performance; hence, performance will vary from system to system. Many old systems were built to extremely loose specifications on noise and cross-modulation and have serious reflection problems because of the use of unmatched subscriber taps. Fortunately, systems in large cities and new buildings are much newer and are required to meet more exacting standards.

Even with these systems, the construction norms are still those which satisfy casual TV viewers, not data users. Thus, loose connections cause intermittant transmission conditions, and momentary "disconnects." These would cause only minor "flashes" on a TV picture but constitute major data dropouts in a high speed data circuit. Finally, systems may be inadequately tested, and hence, in some parts of a CATV System noise and cross-modulation levels may not meet written system specifications. The limiting factor in determining the performance of the system will not be Gaussian noise interference, but a number of practical factors which provide interference, generally categorized as "impluse noise." These factors are difficult to characterize and include phenomena such as loose connections, cracked cable sheaths, and R-F leaks.

Two factors dominate the specification of any data transmission mode on a CATV System.

a. That data is sharing a transmission medium with video signals.

b. That there will be a large number of users.

9.3.1 Interface with CATV System

Two Way Options

The data transmission system must be readily adaptable to a wide variety of existing CATV System designs and two-way options.

Data Rates

The data signals must not cause visible interference with video signals.

Installation

If auxiliary data equipment is to be added to the CATV System, it must satisfy the following requirements:

- It can be installed with only minor changes in the CATV System.
- It need be installed in only a small number of locations.
- It can be installed rapidly in early hours of the morning to prevent interference with TV service.

Low Cost

To maximize the marginal utility of data distribution over the CATV System, any equipment introduced must be inexpensive.

9.3.2 Interface With Population

Population Density Variations

Standard transmission configuration options must be available for systems of various sizes, population densities and percent of active users. Because of the huge number of potential users, all terminal equipment must be simple and inexpensive.

Unsophisticated Users

To minimize user interaction with the system operating mode, all terminal equipment must be the same for each location; it must use the same frequencies and data rates; and it must have no options for equipment modification by the user.

The MITRE Corporation has patented a system called MITRIX which meets all the above specifications [58] and has many other excellent features. Some of the disadvantages of polling for terminal-oriented networks [50] are the synchronization delays [32] and the large amount of channel bandwidth occupied by simply polling 100,000 subscribers. MITRIX overcomes both these problems by using a time division multiple access scheme. Furthermore, since the number of time slots per second is dynamically assigned to a subscriber, the system also avoids the wasted bandwidth in allocating fixed frequency bands (FDM) or fixed time-slots (TDM) to subscribers who are active for only small amounts of time. The users interface unit requests a certain number of time slots within frame periods and these are then allocated by a Computer Digital Interface Unit; a DEC PDP-15. The system is highly flexible in structure, efficient and inexpensive.

9.3.3 Other Considerations

However, there are still some tradeoffs involved and for large systems improvements are still in the offing. In particular, we are still faced with the problem that if a subscriber logs in at a terminal and makes a request for a certain data rate then he holds those time slots until he logs out whether he is actively typing or thinking and not typing. For small systems or systems with a small number of users, this may be an acceptable inefficiency in bandwidth use. But for systems like the suburban Boston system it may not be acceptable. As we shall see the factors involved are the available bandwidth, the average ratio of active user time to inactive user time in a logged-in period, and the average number of active users. The alternative which makes more efficient use of bandwidth for high peak to average data rates is the random access packed multiplexing method previously derived and applied to the satellite channel. With packets, terminals seize the channel only when they are active. Hence, more users can be accommodated. Furthermore, reservations of channels are not required. As we have seen the random access feature results in a channel availability of 1/2e of the band or 1/e for a slotted system. This is effective if the average peak to average data rate is greater than the number 2e or e, respectively.

9.4 Number of Active Terminals for Random Access Packet Sample System

We have seen in Chapter 5 that for an unslotted random access packet channel the maximum number of active terminals k_{max} is given by $(2e\lambda\tau)^{-1}$. If we let d be the pulse duration and let γ be the number of pulses per packet, then k_{max} is $(2e\lambda\gamma d)^{-1}$ where $\lambda\gamma$ has the dimension of pulses/second per terminal. For a two level FSK system, this is the same as bits/second per terminal. In Figure 9.4 we plot the maximum number of active terminals versus $\lambda\gamma$ for the systems in Table 9.2 using the above equations. The curves labeled A, B,C, and D correspond to the slotted system in Table 9.2 labeled A,B,C, and D. The lines labeled A', B', C', and D', are for the corresponding unslotted systems.

We can now examine Figure 9.4 to determine system performance under some typical data transmission requirements. For a data rate of 40 bits/second per terminal, a single trunk can handle 4000 terminals with a slotted system (1 Megabit/sec) with an error rate of 10^{-22} at a signal to noise ratio degraded by 20db. The average number of TV sets per trunk in the suburban Boston system is approximately 27,000. Hence, the simplest

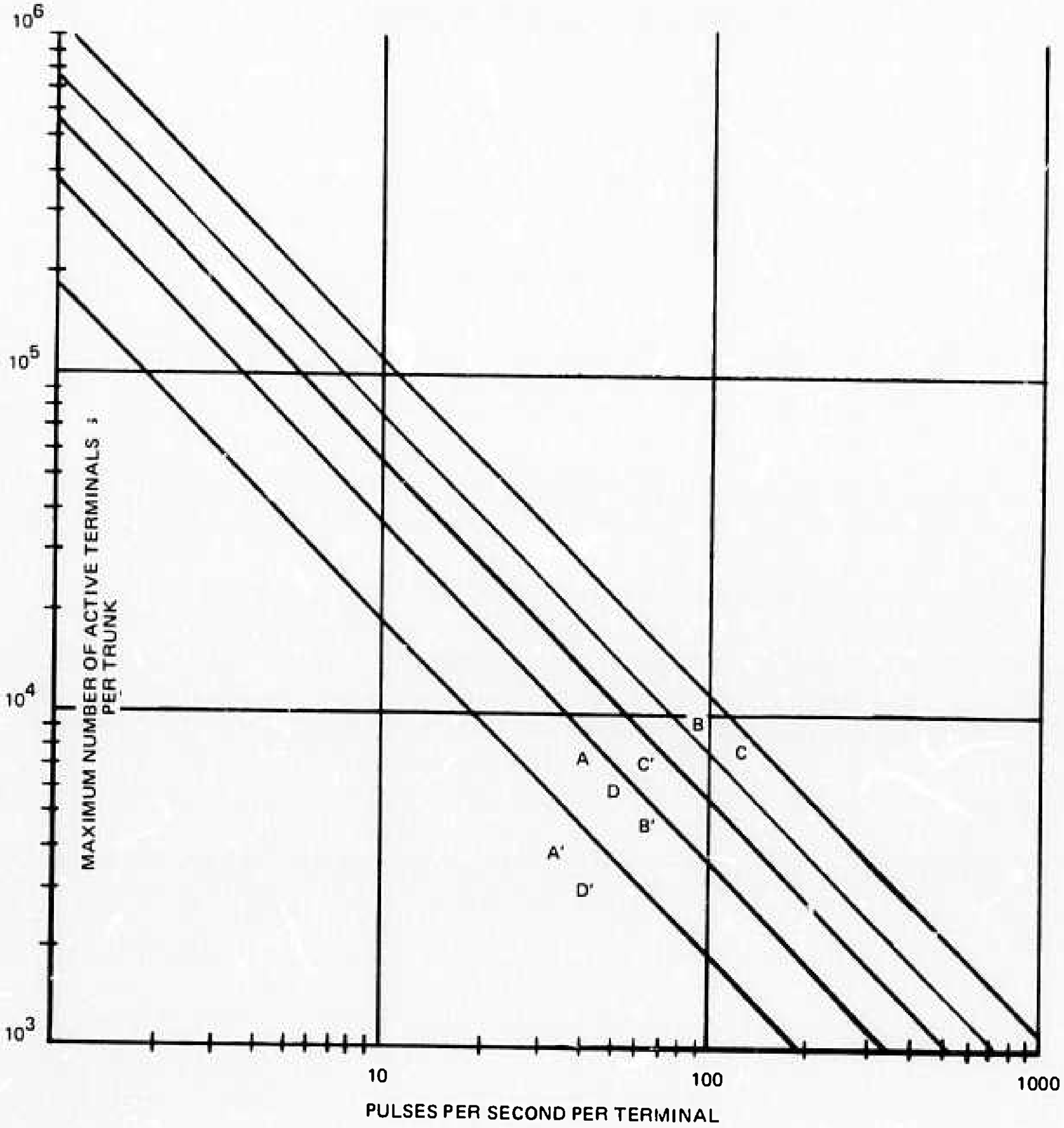


Figure 9.4: Performance of Random Access Packet Cable System

modulation scheme will handle one third of all terminals as active terminals. At 100 Kbits/second, the system will handle 900 active terminals.

The reason for considering a 100 Kbits/second channel instead of 1 Megabit/second channel is that it allows more adaptability for local point-to-point traffic with the addition of standard digital devices such as concentrators or routers. Of course, if this is not required, the 1 Megabit channels can be used.

Table 9.2: Number of Active Users Per Trunk

Type of System Data Rate	Slotted System	Unslotted System
1 Megabit/sec.	9,000	4,500
100 Kilobit/sec.	900	450

We will use the terminology of the cable TV industry in describing the direction of signal flow. Signals traveling from the head end toward terminals will be said to be directed in the "forward" direction on a "forward" link and signals traveling from terminals toward the head end will be said to be directed in a "reverse" direction on a "reverse" link. A convenient synonym for "forward" will be "downstream," and for "reverse" wil be "upstream." To install the device to be described in the forward and reverse channels simple duplex and triplex filters can be used. The devices are illustrated schematically in Figure 9.5.

9.4.1 Carrier Frequency Conversion

In the simplest version of a data system, two carrier frequencies are used; one for forward transmission from the head end to the terminals, and one for reverse transmission from the terminals to head end. Let us call these angular frequencies ω_f and ω_r respectively. The next simplest option is to use frequency converters at a small selected set of points in the system. In the forward direction, the converter converts from ω'_f to ω_f and in the reverse direction, it converts from ω_r to ω'_r. The net result is that the terminals still receive and transmit at the frequencies ω_f and ω_r. However, in the trunk between the converters and the head end, there are four frequencies in use, ω_r, ω'_r, ω_f and ω'_f so that in these trunks twice the traffic can be handled.

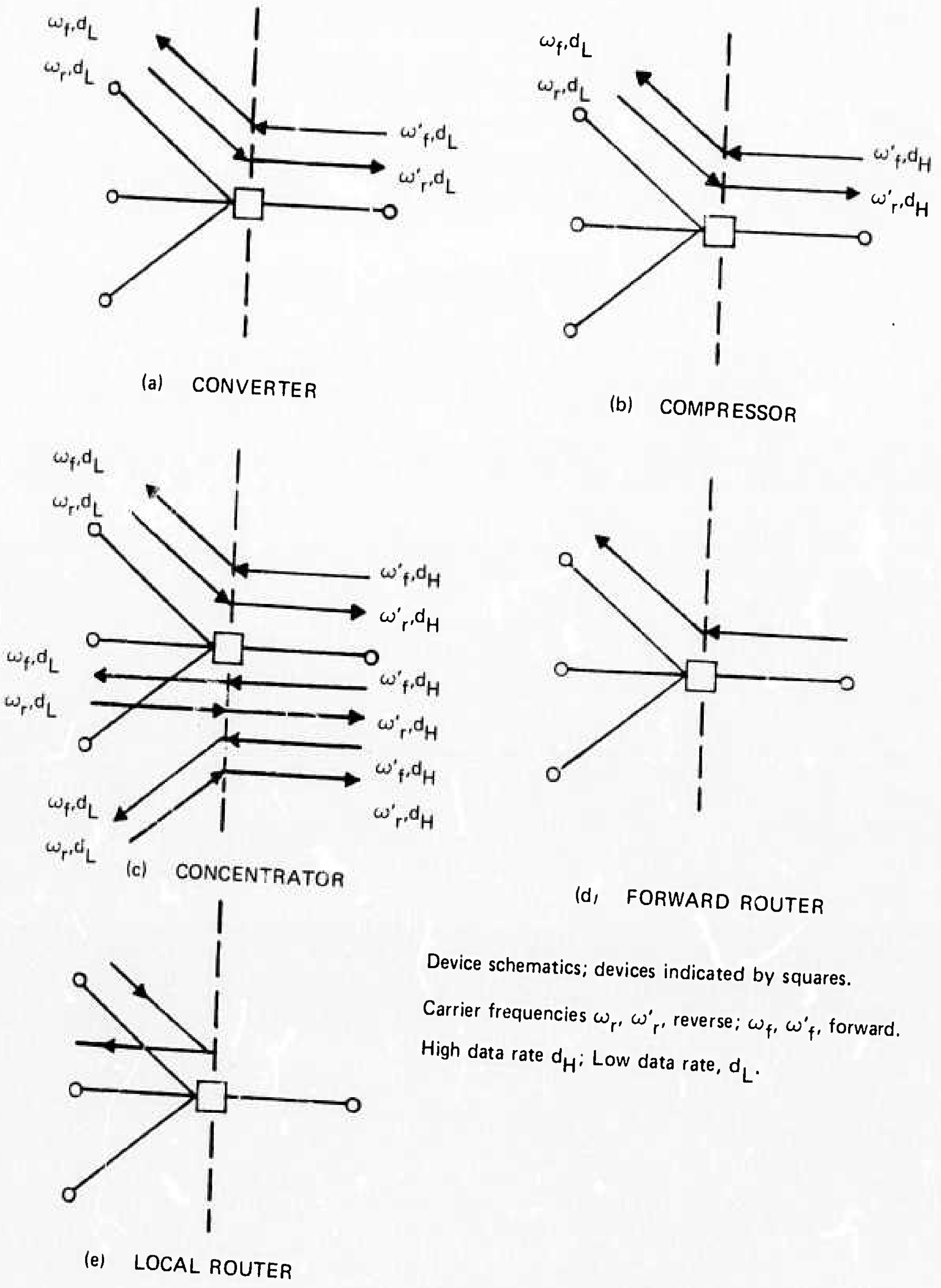
The advantages of this scheme are:

a. All terminals are identical.

b. The capacity of the system is increased since two channels are available in each direction for heavy traffic sections of the cable.

9.4.2 Routing

There is no requirement for routing since a basic premise is that all receivers listen to all messages that reach them and merely select the ones addressed to them. Nevertheless, we will consider the addition of some primitive low cost routing schemes and qualitatively indicate their effect on system capacity.



(a) CONVERTER

(b) COMPRESSOR

(c) CONCENTRATOR

(d) FORWARD ROUTER

(e) LOCAL ROUTER

Device schematics; devices indicated by squares.

Carrier frequencies ω_r, ω'_r, reverse; ω_f, ω'_f, forward.

High data rate d_H; Low data rate, d_L.

Figure 9.5: Concentration Alternatives for Packet Cable System

In the central transmission mode there is no routing needed in the reverse direction since all messages reach the head end along the unique paths from the originating terminal. In the forward direction, the signals at frequency ω_f are blocked by filters at the converters and yields a simple form of routing. In fact, at any section of trunk not requiring signals at ω'_f filters can be added to block ω'_f. Adding these filters does not increase system capacity but may be useful if the frequency ω'_f can be used for local signaling when not being used for data transmission.

At any junction containing a converter, digital routers *must* be added to send messages at ω_f down the trunk to which they are addressed rather than all trunks. Such a router may be added at any other point in the system as well. This is called *"forward routing"* and can increase system capacity. Forward routing requires a digital router which can read and interpret message addresses.

Let us now consider these system options in the presence of local traffic. The option of frequency conversion is unaffected and performs in exactly the same manner as in the central transmission mode. However, an extra routing option is available for local transmission. In particular, if two terminals are on the same trunk, then the message between them can be intercepted and routed at a routing station rather than travel all the way to the head end. Such routing is called *"local routing."* Local routing reduces the traffic on the main trunk.

9.4.3 Compression

As the next more complicated option, the data rate as well as the carrier frequency is changed at a converter, i.e., a compressor can be used. The advantages of this arrangement are:

a. All terminals operate at low data rate.

b. On heavily used lines near the head end, a higher data rate, say one megabit/second, can be used to increase the number of potential active users or decrease the delay.

c. Even though a section of the trunk can carry high data rate traffic at carrier frequencies ω'_r and ω'_f, other users can still use the system at the low data rates at ω_r and ω_f. Thus, the number of compressors required is small.

9.4.4 Concentration

Finally, the compressor at junctions may be replaced by a concentrator. That is, messages arriving simultaneously on two or more links in the reverse direction are buffered and sent out sequentially at the higher data rate. This essentially makes the system downstream from the concentrator appear to operate at the higher data rate and hence increases the system capacity even further.

9.4.5 Frequency Division Multiplexing

In case the data rate is limited by the head end mini-computer, an available option is to frequency division multiplex several 100 Kbits/second channels, each of which is processed by a separate head end minicomputer.

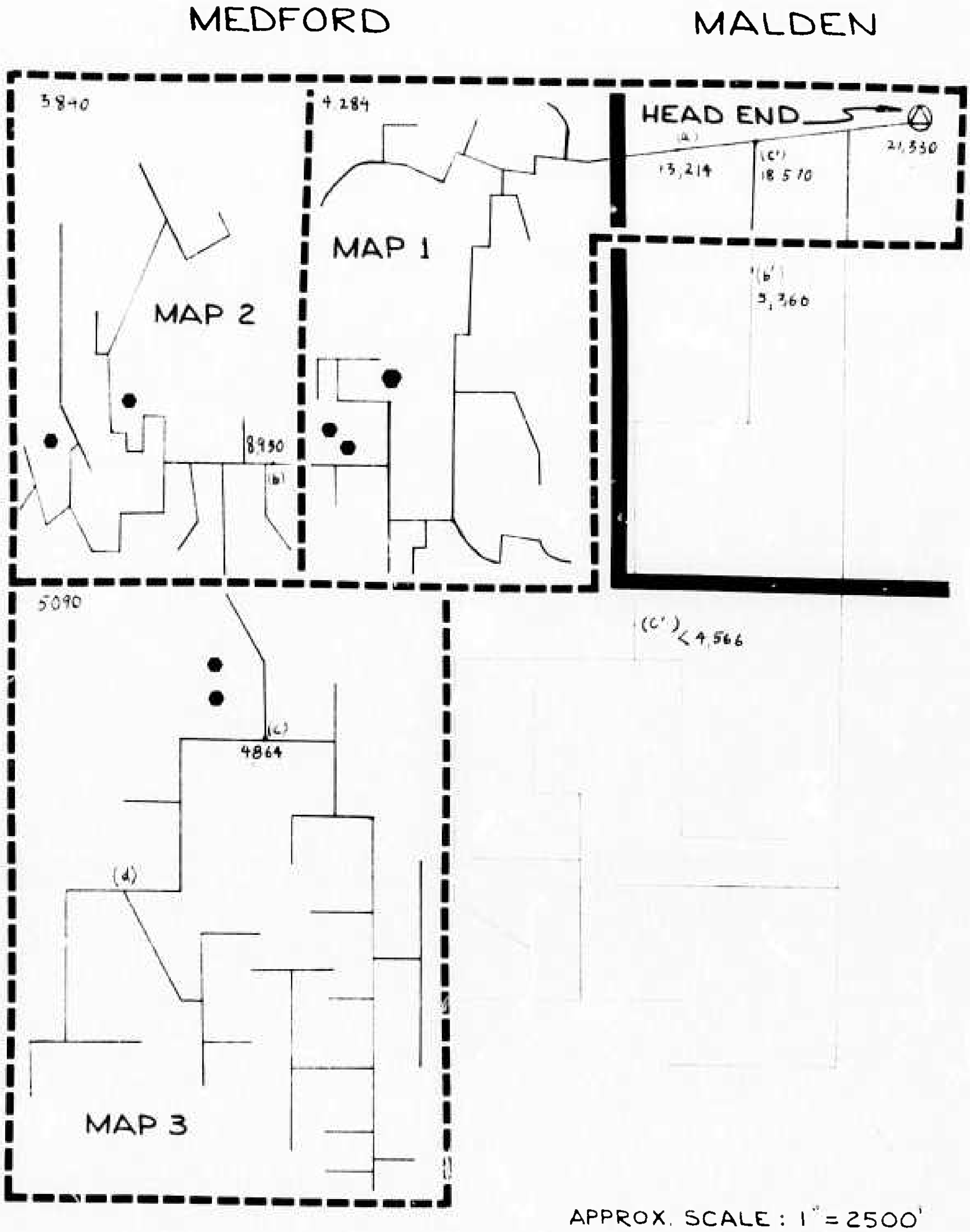
The assignment of these options in an optimal fashion requires detailed expressions for the traffic in the links. Formulae for the traffic in links using combinations of the various digital devices are fairly obvious in detail although rather tedious to present in generality.

9.5 Random Access Packet Designs For Sample CATV System

We will illustrate the usefulness of the various options and devices we have considered, for example, in adapting to a low data rate channel system of 100 Kbps. We apply the devices to a design of a random access packet data system for Medford, Massachusetts, a section of the Suburban Boston CATV complex. In Figure 9.6, a branch of the trunk is drawn for Medford. The triangles represent bridger amplifiers. These amplifiers feed into feeder cable and extender amplifiers with customer taps and drop lines emanating from the feeder cable. The feeder backer arrangement previously mentioned is used. The feeder system emanating from a given bridger amplifier is called a cluster. The number next to each amplifier gives the number of terminals in the cluster associated with that amplifier. The average number of terminals per cluster is 137 with complete coverage of all homes.

We focus our attention on one trunk in the Medford area. We assume that the average traffic per terminal is 40 bits/sec. We assume most conservatively that in the design the data from the terminals is at 100 Kbps; the upconverted rate is 1 Megabit/sec. We have already obtained the results in Table 9.2 showing the number of active terminals that can be supported on a trunk at each data rate.

Since we have not presented relative costs of converters, concentrators, multiplexers and routers, we are not optimizing the design. We are merely presenting feasible designs to demonstrate the wide range and flexibility achieved by combinations of a few devices. The designs are easily described by simply indicating the location of the various devices on the map in terms of an alphabetic label on the map. To aid in visualizing the design, the number in the rectangle beside the letter (on the key maps) indicates the population downstream from that point. The designs are as shown in Table 9.3.



APPROX. SCALE : 1" = 2500'

Figure 9.6a: Key Map

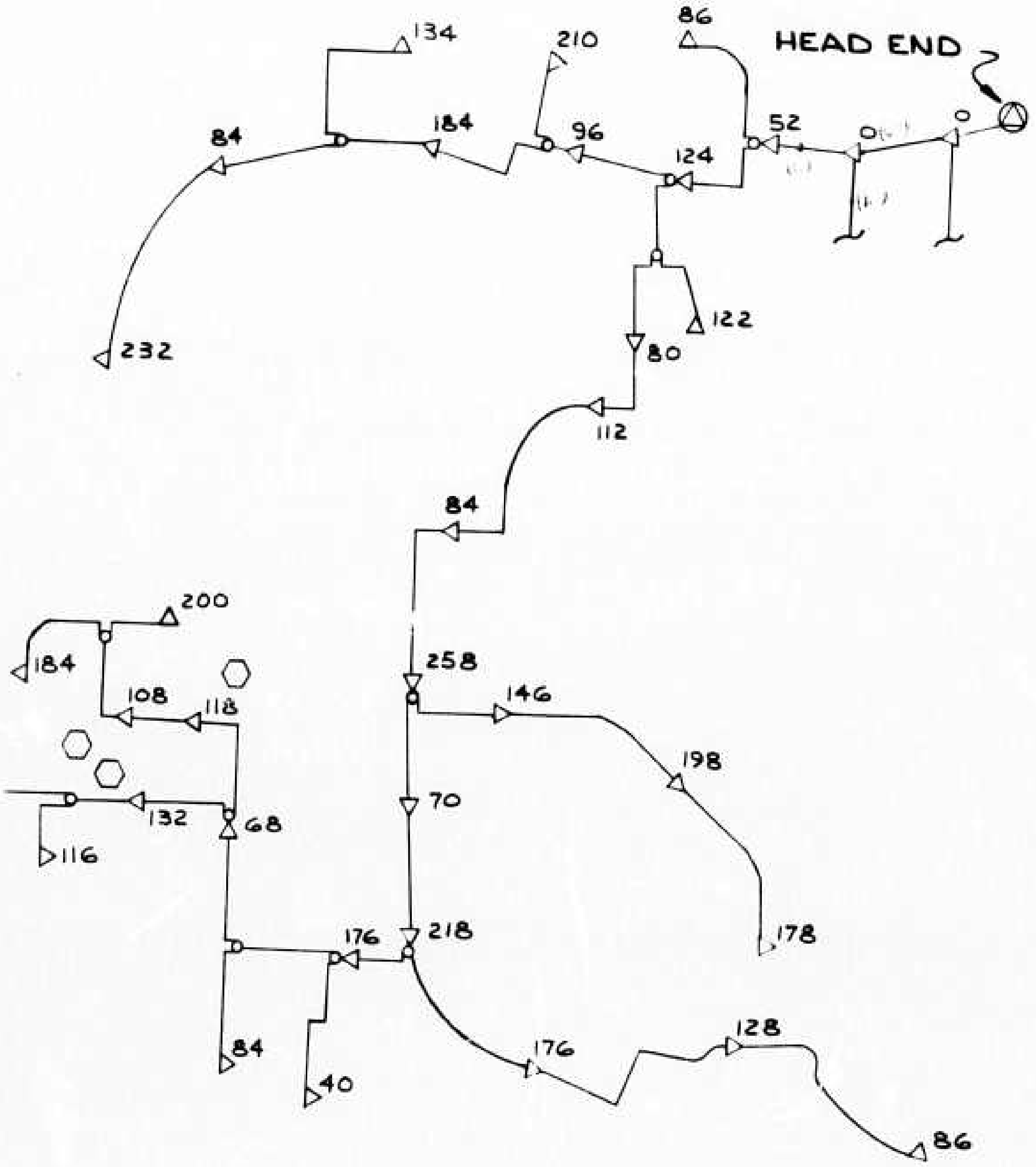


Figure 9.6b: Map 1

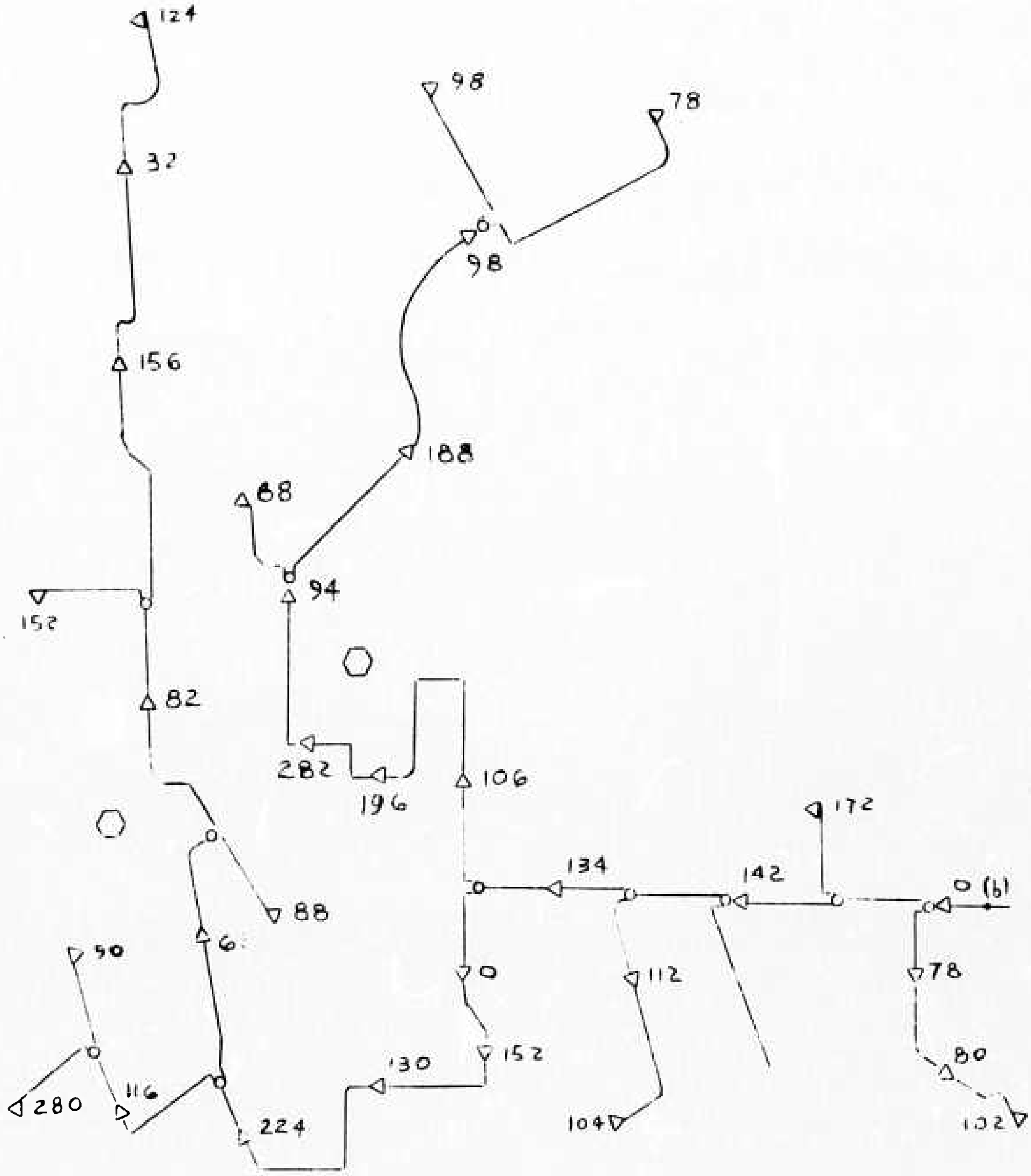


Figure 9.6c: Map 2

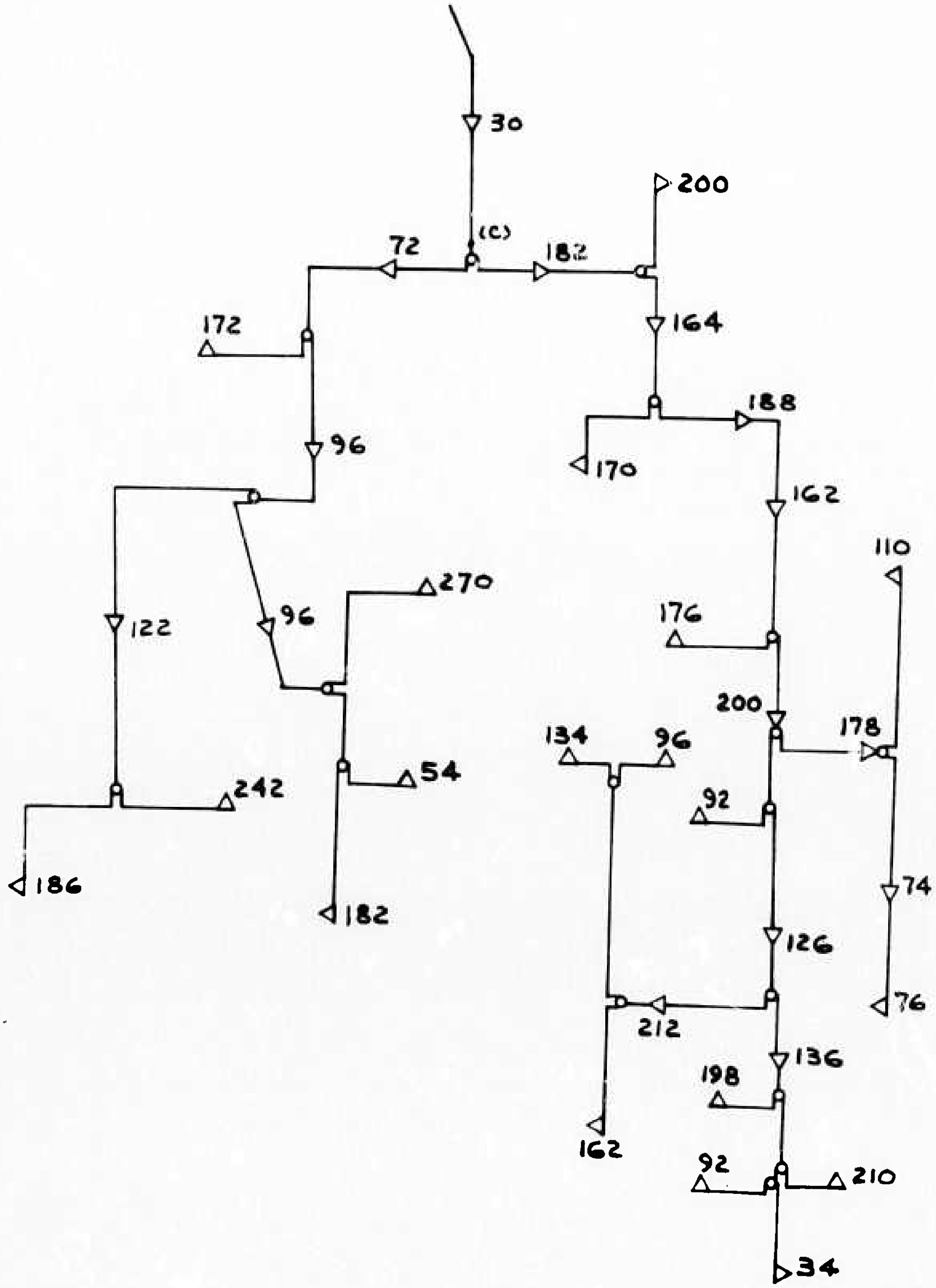


Figure 9.6d: Map 3

Table 9.3 Feasible Designs for Central Transmission Mode At 100 Kbps Data Rate

% Active Terminals	Slotted	Unslotted
1%	No devices	No devices
3%	No devices	Converter at (a)
10%	Compressors at (b) & (b′)	Compressors at (c′′) and Concentrators at (b), (c), and (c′)
15%	Compressors at (c′′) and Concentrators at (b), (c) and (c′)	

Chapter 10
PACKET RADIO NETWORKS FOR LOCAL ACCESS–INTRODUCTION

10.1 Network Overview

The main features which distinguish the Packet Radio System from a point-to-point packet switching system (such as the ARPANET) are: (i) devices in the system transmit packets by using a random access scheme, and (ii) devices broadcast so that packets can be transmitted to several devices simultaneously, and/or several packets can be simultaneously received by a receiver due to independent transmissions of several devices. These features have a major impact on practically every aspect of network considerations.

There are three basic functional components of the Packet Radio System: the Packet Radio Terminal, the Packet Radio Station, and the Packet Radio Repeater. (See Figure 10.1.) Packet Radio Terminals will be of various types, including personal digital terminals, TTY-like devices, unattended sensors, small computers, display printers, and position location devices.

In some applications the Packet Radio Station will be the interface component between the broadcast system and a point-to-point network. As such it will have broadcast channels into the Packet Radio System and Link channels into the point-to-point network. In addition, it will perform accounting, buffering, directory, and routing functions for the overall system.

The basic function of the Packet Radio Repeater is to extend the effective range of the terminals and the stations, especially in remote areas of low traffic, and thereby increase the average ratio of terminals to stations. A more detailed discussion of the network hardware functions can be found in Section 10.2.

The devices (repeaters, stations, and terminals) of the Packet Radio System communicate in a broadcast mode using a varient of the Aloha random access method [1].

Stations will be allocated on the basis of traffic. Thus, to first approximation, we can think of partitioning the area to be covered into regions of equal traffic and allocate one station for each region. In regions of low traffic density, the station may not be in "line of sight" of all the terminals in the region; hence repeaters are used to relay the traffic to the station. Thus, repeaters correspond to a geographical partition of the area into sections

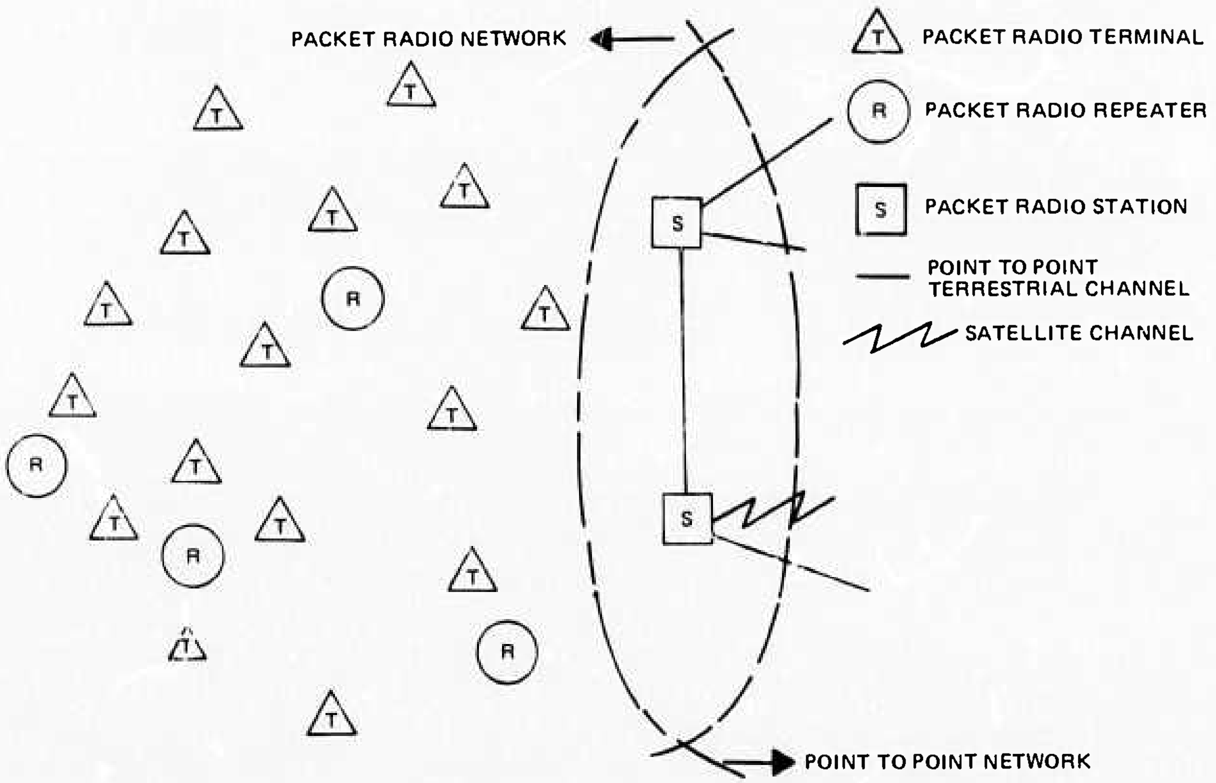


Figure 10.1: Packet Radio System

small enough so that each terminal can communicate with a repeater and its messages be relayed by repeaters to a station.

In areas of high traffic, such as urban areas, repeaters may not be needed: in fact, the problem may be that a station can communicate with more terminals than it can handle. Broadcast of data in urban areas is also complicated by multipath interference [52]. The rapidly expanding Cable Television Systems within urban areas offer an attractive alternative to over-the-air broadcasting, except for mobile users who must use broadcast techniques. As we have seen, the same general packet radio concepts can be applied to broadband Cable Systems.

10.2 Network Elements

10.2.1 Nodes

In this section we discuss the devices' functional capabilities which are necessary for *communication* in the Packet Radio network. Functional requirements of these elements not directly related to communication are not discussed.

Terminals

There are two categories of terminals; (a) those which usually await a response to a message they transmit (e.g., manually held radio terminals, small computers), and (b) those which do not require such responses or acknowledgements (e.g., unattended sensors, position indicators). Some terminals in the former category will usually send and/or receive several packets in one message.

Necessary or desirable communication capabilities of a terminal:

a. Ability to identify whether the packet is addressed to its ID.

b. Calculation of packet checksum.

c. Capabilities related to packet routing such as; retransmitting packets when acknowledgements are not received, recording and using a specific ID of a repeater and/or station to be used for other packets of the same message, counting the number of retransmissions.

d. Capabilities related to the response to previously determined types of error.

e. For unattended terminals, capabilities by which a centralized control or a station will be able to identify whether the terminal is operative or dead.

Repeaters

Functional capabilities for repeaters include:

a. Calculating packet checksum.

b. Packet storage and retransmission.

c. Capabilities by which a station can determine whether a particular repeater (or any repeater in a particular area) is operative or dead.

d. Capabilities 1, 3, and 4 of terminals.

e. Capabilities, dependent on the routing strategy, for calculating the most efficient next repeater on a transmission path to a station or to a terminal.

Stations

Among the stations' functional capabilities are:

a. A directory of terminals and repeaters in its region.

b. Operations necessary to convert packets from the Packet Radio System into packets used in the point-to-point network and conversely.

c. Storage buffers for packets received from terminals and for packets to be transmitted to terminals.

d. Storage for character position information for active terminals which do not have this capability.

e. Accounting capabilities.

f. Capabilities related to routing, flow control, and network management.

10.2.2 Channels

Communication between devices is by broadcast, using a variant of the ALOHA random access method. Many aspects of the broadcast channel are of peripheral interest in the network design of the system; however, some factors are crucial for determining the behavior of the network. By ALOHA transmission we mean the use of a shared channel which is randomly accessed by more than one user. In the simplest case users transit equal size packets, each using a data rate equal to the channel data rate several modes of operation are possible. The two simplest are: a non-slotted (asynchronous) mode in which users can access the channel at any time, and a slotted (synchronous) mode in which users can access the channel only at the beginning of a slot of time duration equal to a packet transmission time. In the latter case, a form of synchronization is required since each user must determine the beginning time of each slot. The following theoretical results assume that, if two or more packets overlap, none is correctly received, and each must be retransmitted. Such a system is called a system without capture.

The simplest analytic results assume that there are an infinite number of users and that the point process of packet origination and the point process of packet originations plus retransmissions are Poisson with mean S and G, respectively; constant transmission time T for each packet is also assumed. Then, if a packet begins at some random time, the probability that it is correctly received (no overlapping, collision or conflict) in the nonslotted case is e^{-2GT}. The reception rate, equal to the origination rate (assuming that colliding packets are retransmitted until correctly received) is $S = Ge^{-2GT}$. The effective channel utilization is $ST = GTe^{-2GT}$, and the maximum utilization is $\text{Max}(ST) = 1/2e$. For the slotted case, the probability of collision is e^{-GT} which leads to $1/e$ as the maximum utilization. GT, the channel traffic is equal to 1/2 and 1 at a maximum effective utilization for the non-slotted and slotted case, respectively [1].

In the original ALOHA system, implemented at the University of Hawaii [2], a central station communicates with several remote sites. The system contains two channels—one for station-to-site traffic and the second for site-to-station traffic. This has several advantages for the ALOHA system. First, the station broadcasts continuously to furnish synchronization between all sites. Second, station-to-site traffic is coordinated by the station so that messages from the station do not collide with one another. Thus, if the traffic from the station has a separate channel from the reverse traffic, retransmissions are substantially reduced. Allocating separate channels for inbound and outbound station traffic is not as attractive when repeaters and multiple stations are introduced. This channel allocation problem is presently under investigation. Channel improvements also appear to be possible by using Spread Spectrum Coding, which offers the possibility of time capture. Competing packets arriving during the transmission time of the first may be ignored if their signal strength is not too great. When the transmitters are widely distributed, geometric or power capture is also possible [46]. With or without spread spectrum a competing signal which is much weaker (further away) than the desired signal will not interfere. Both types of capture can give rise to performance superior to that predicted by the simple unslotted ALOHA model. However, capture biases against more distant transmitters since the probability of a successful transmission to the station decreases as the distance from the station increases. Hence, it results in the increase in the number of retransmissions and consequently in the delay.

Chapter 11
PACKET RADIO NETWORK TOPOLOGY

11.1 General Considerations

Many factors affect the location of repeaters and stations. Simple consideration of repeaters as area covers and stations as traffic covers neglects interactions between the two types of devices.

Factors affecting the location of repeaters and stations in addition to range and traffic are:

a. **Logistics:** Some locations for repeaters may be preferable to others because of greater accessibility or more readily available power, eliminating the need for batteries (e.g., on telephone poles or near power lines).

b. **Reliability and Redundancy:** For many reasons, redundant repeaters and stations will be required. Since repeaters in remote areas will operate on batteries, it will be necessary to have sufficient redundancy so they need not be replaced immediately. Stations and repeaters will have intermittent and catastrophic failures for which backup is required. Extra repeaters are needed when line of sight to the primary repeater is locally blocked.

When a single channel is operated in an unslotted ALOHA random access mode, no more than 1/2e of the bandwidth can be effectively utilized, as discussed in the previous section. However, additional traffic is generated by repeaters, and conflicts created by transmissions between adjacent stations. Some sources of retransmissions are:

a. For reliability, several repeaters or stations must be within range of each terminal. If the repeaters retransmit every packet they receive, one message can generate an exponentially growing number of relayed messages. To prevent one message from saturating the network, traffic control is required. *The discipline chosen and its efficiency will probably be the single most important system factor affecting system performance.* Two types of undesirable routing through the repeaters can occur. First a message can circulate endlessly among the same group of repeaters if not controlled. Second, even if no message is propagated endlessly, a message can be propagated to a geometrically increasing number of new repeaters in a large network.

b. For system reliability, more than one station must be able to transmit via repeaters to each terminal. Thus, there can be conflicts between adjacent stations which reduces the useable bandwidth and also introduces coordination and routing problems.

c. In general, there will be many routes between any given terminal and any given station. Consequently, more conflicts can result than would be the case if the terminals communicated with a station.

11.2 Device Location

To provide line-of-sight coverage of an area where mobile terminals or fixed terminals are transmitting by radio from unspecified locations we must locate *repeaters* so that any such terminal will be in line-of-sight of repeaters and that there be reliable connections between every pair of terminals (and repeaters). More precisely, we wish to minimize the installation cost and maintenance cost of the repeaters subject to a constraint on the reliability of service.

In general, determining if line-of-sight microwave transmission between two points is possible involves taking into account many factors including wave-length (Fresnel zones), weather conditions (effective earth radius), antenna design, height, topography, etc. Nevertheless there are methods for making such calculations [42]. In this section, we describe methods for using the results of these determinations to choose good locations for the repeaters.

It is impractical to consider all possible locations of repeaters and terminals, which theoretically are infinite in number. We limit outselves to a finite set R of possible repeater locations and a finite set T of possible terminal locations. How the set R and T are chosen will be of great computational importance and will probably be chosen adaptively. But for the time being, we assume R and T known and fixed.

The principal and immediate interest is in an appropriate mathematical model of the situation and some indications on how to solve the problem. The first problem is the proper choice of reliability measure or grade of service. We assume that the radio network is for local distribution-collection of terminal traffic with rates small compared to the channel capacity so that throughput capacity is not a constraint. That is, if any path through the network exists for a given pair of terminals we assume there is sufficient capacity for traffic between them. Possible measures of network reliability that have proved useful in the analysis of communication network [55] are the probability that all terminal pairs can communicate and the average fraction of terminal pairs which can communicate. However, for network *synthesis* as distinguished from *analysis* these measures appear too difficult both from computational and data collection points of view. This suggests the "deterministic" requirement that there exist k node disjoint paths between every terminal pair. This guarantees that at least k repeaters or line-of-sight links must fail before any terminal pair is disconnected. Let the cost of a repeater at location $r \in R$ be $c(r)$ and $c(R^{\circ}) = \Sigma \{ c(r) \mid r \in R^{\circ} \}$ where $R^{\circ} \subset R$. Then, we can formulate:

Problem I

Find $R^* \subset R$ minimizing $c(R^*)$ subject to the constraint that for all $t \in T$ and $r \in R^*$ there exist k node disjoint paths from t to r.

One might demand only that there be k node disjoint paths between every pair of terminals instead of between each terminal-repeater pair, but we are assuming that communication always takes place through a "station" which could be any of the repeaters. The analysis of the terminal to terminal model is similar in any case.

The constraint can be broken into two parts:

a. k-fold set covering: The repeaters must be located so that at least k of them are in line-of-sight with each terminal, and

b. k redundancy: Between each pair of *repeaters* there must be k node disjoint paths.

Because the repeaters will have substantially greater range than terminals, the first aspect of the constraint will ordinarily be dominating. Moreover, it can be shown that the problem of minimizing costs of repeater locations subject to k-cover constraints is mathematically equivalent to 1-covering.

The 1-cover problem is the classical set covering problem. Extensive research has been and is being done on this problem, but there is good evidence—empirical [26] and theoretical—that the problem is intrinsically difficult.

Given the limited success to be expected from exact algorithms in solving large scale problems, we have been led to consider heuristic methods to find good solutions to the k-cover (of terminals by repeaters) problem which is typically large scale. It is intuitively appealing to consider a *terminal* as particularly critical if it is adjacent to few repeaters. (In the extreme cases, if a terminal has fewer than k adjacent repeaters, the problem is infeasible and if it has exactly k adjacent repeaters, all of them must be chosen for any feasible solution). Similarly, a *repeater* is desirable if it is adjacent to a large number of terminals, especially if the terminals are highly critical. The heuristic algorithms systemize these intuitive notions in the search of a "good" solution.

The size of test problems solved varies from problems with as few as 5 repeaters and 5 terminals to problems with as many as 400 repeaters and 400 terminals. Roughly speaking, the computation time was directly proportional to the size of the incidence matrix and the cover multiple required. The computer used is a PDP-10 (time sharing). The larger problems (400 repeaters, 400 terminals, 2-cover) were solved in 70 sec or less. The time, as may be expected, is dependent on the density of 1's in the incidence matrix. Thus, the maximum time recorded arose from terminals—repeaters configuration where each repeater covers many repeaters. The running time is of the order of $|T| \times |R|^2$ where $|T|$ and $|R|$ are the number of terminals and repeaters respectively.

We ran a number of problems with the heuristic code and for comparison with the Ophelie mixed integer programming code running on a CDC 6600 computer. The Ophelie code uses the branch-and-bound method. In the case of very simple problems (8 repeaters, 9 terminals, 2-cover) there was essentially no difference in running time (presumably most of the time, less than .5 sec, was spent in setting up the problem). Running experience with the Steiner triples' problem described in the next section, yields a ratio of 500 to 1 between the Ophelie time and the heuristic code time when solving the smaller problem A_{27} (117 terminals, 30 repeaters, 1-cover, and no comparison is available for the larger problem A_{45} (330 terminals, 45 repeaters) since for example, the MPSX code failed to reach a solution in more than one half hour on an IBM 360-91.*

Comparison in running time is naturally not completely valid, since most of the computation time in the Ophelie code can be spent just checking if a given solution is optimal. The heuristic method does not try to check the optimality of its solution However, in general, results with the heuristic code have been extremely good. When the heuristic solution deviated from the optimal solution, the problem usually involved numerous tries for the maximum $\omega_i^{(\ell)} \ell = 1, 2, 3, 4$ such as in the Steiner triples' problems. In all problems that were generated to resemble the packet radio terminal-repeater problem, the heuristic algorithm reached the optimal solution (in those problems for which we are able to determine the optimal solution).

[23] report on two covering problems which they characterize as computationally difficult. Each problem is defined by the incidence matrix of a Steiner triple system. The first problem, labelled A_{27} is a 1-cover problem with 117 terminals and 30 repeaters. The second problem, labelled A_{45}, has 330 terminals and 45 repeaters and is also a 1-cover problem. Data for both problems can be found on pages 9 and 10 of [23]. The problems are considered to be difficult because of the large number of verifications (branching in branch-and-bound, costs in cutting methods) required to establish that a given solution is in fact optimal.

The heuristics developed are dependent on the order in which the repeaters are presented to the algorithm. One hundred random permutations of the repeater ordering was tried for each of the two Steiner problems. In each case, the heuristic obtained the optimal solution for the smaller problem and for the larger problem, three solutions out of the 100 equaled the conjectured optimal solution of 30 repeaters. The remaining 97 were within

*Private Communication, [23].

3% of the optimals. On the other hand, we have constructed other artificial problems on which the heuristic performs abysmally.

Another test problem used to compare various techniques to solve the k-covering problem was generated by using real data obtained from a topographical map for the region of Palo Alto. This part of the U.S. was selected because it contained many interesting topographical attributes: a flat terrain (salt flats, the region surrounding the Bayshore Freeway), an urban center (Palo Alto and neighboring communities) on slightly sloping terrain and finally a hilly region (with valleys, small plateaus, etc.). Moreover, at this time, it appears that a reduced scale experiment of a packet radio network will be installed in the Palo Alto area.

LOS Computation

To determine if a terminal at location j can be seen from a repeater at location k, we proceeded as follows. It was assumed that if nc particular high construction (building, water tower, etc.) was available to install the repeater's antenna, it would be installed at 30 feet above the ground level (making use of a tree, telephone pole, etc.). The terminals were assumed to be 5 feet above ground level. The points were said to be in LOS if the first *Fresnel Zone* associated with transmission between these two points was free of any obstacle (see Figure 11.1).

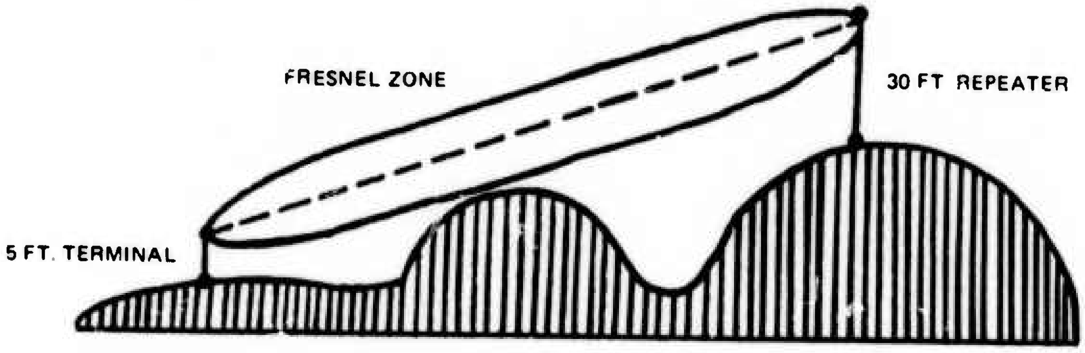


Figure 11.1: Relationship Between Terrain and Antenna Heights for Fresnel Zone

To compute the Fresnel Zones, we assumed that transmission would occur at 1500 MHz corresponding to a wave length λ = .2m (7.87 in.).

The problem was solved by the heuristic algorithm and by Ophelie. (A rapid analysis of the terminal-repeater adjacency matrix shows that none of the optimal solutions would have been generated if one had used the more simplistic approach of selecting the repeater with highest adjacency degree. Such a selection yields quite different answers requiring a larger number of repeaters).

The optimal solution requires the installation of 14 repeaters (different runs with the heuristic showed that there were, in fact, a number of optimal solutions with 14 repeaters). The total running time for Ophelie was approximately 12 CPU sec excluding set up time. The SETCOV required 3 sec to produce a solution. The relative success of the Ophelie code must, at least in part, be attributed to the fact that the linear programming solution (which is used to initiate the branch-and-bound part of the code) is actually the optimal solution. The input contour configuration and repeater location solution for this sample problem are shown in Figures 11.2 and 11.3, respectively.

11.3 Network Topological Reconfiguration

From the general topological considerations, it is apparent that the routing and flow control algorithms will be the main factor which will determine the efficiency of the Packet Radio System. However, there are two contradictory requirements; reliability considerations advocate that every repeater should be able to transmit to several repeaters; on the other hand, efficiency consideration suggest that one repeater should receive and relay the packet, preferably the repeater along the shortest path to the destination. A sensible solution is to assign to the set of repeaters a structure which will transform the broadcast network to a point-to-point network for routing purposes. The problem is that the connectivity of devices is changing, and therefore, it is necessary to develop algorithms for dynamically changing the network structure (reconfiguration) under certain conditions. Examples of a changing topology are when the network is mobile (e.g., a Packet Radio System for a fleet of ships), drainage of battery power of repeaters placed in inaccessible environments, or when repeaters fail to operate.

In [41], we propose algorithms for dynamically changing the network configuration. It is assumed that every repeater and station have a fixed ID, and that there is a simple routing algorithm however inefficient, which is independent of any network structure. The process contains three steps:

STEP I

Mapping the network connectivity. This is obtained by a process in which stations transmit packets to repeaters, requesting each to respond with a trace packet into which every repeater along the path adds its fixed ID.

STEP II

Determining network structure. The connectivity information obtained above is used to obtain a network structure which has several properties; for example, it enables every packet to be routed along the shortest path (minimum number of hops); it determines the repeaters which are not needed for relaying packets and which should be temporarily disabled.

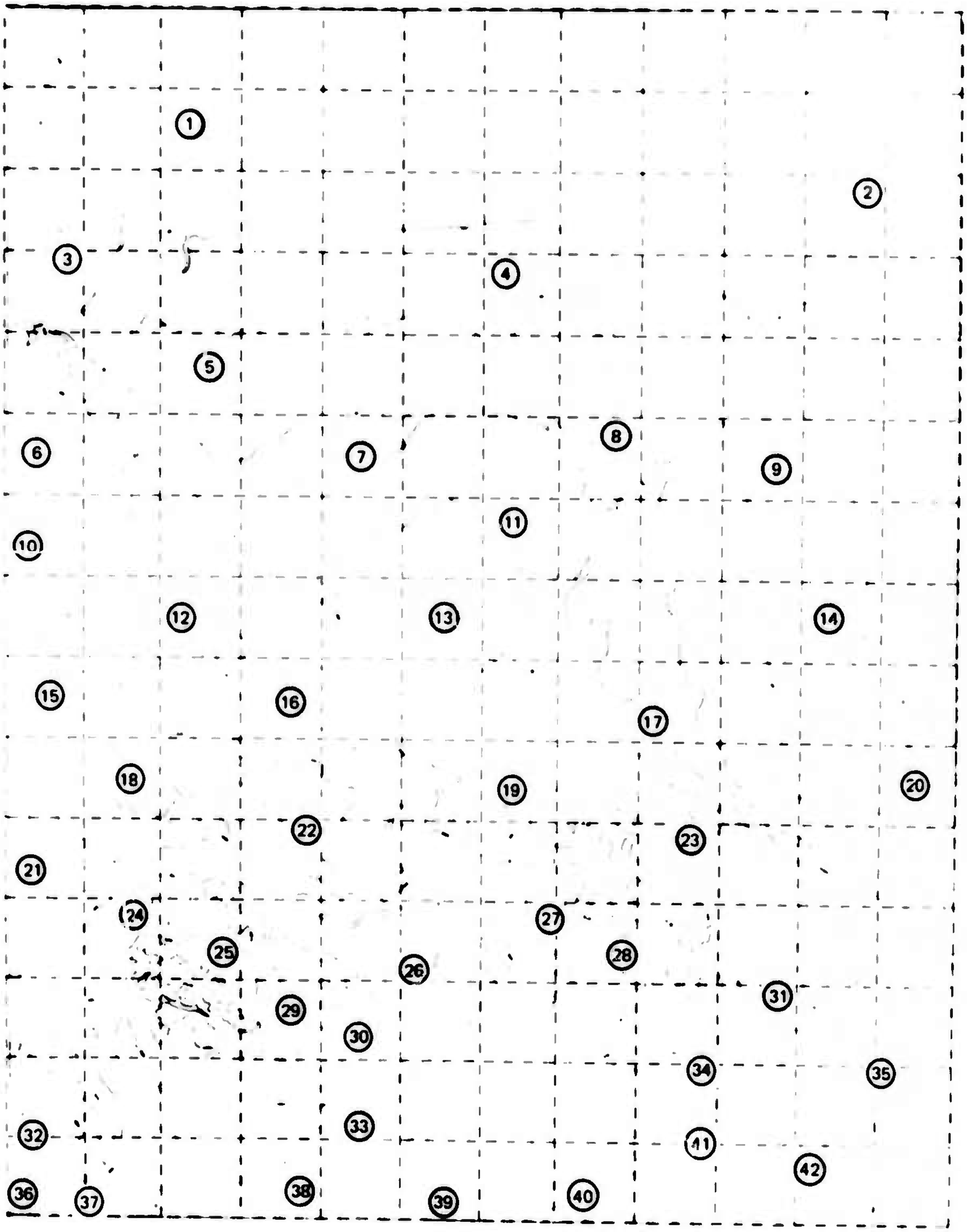


Figure 11.2: Contour Map and Available Repeater Locations for Example

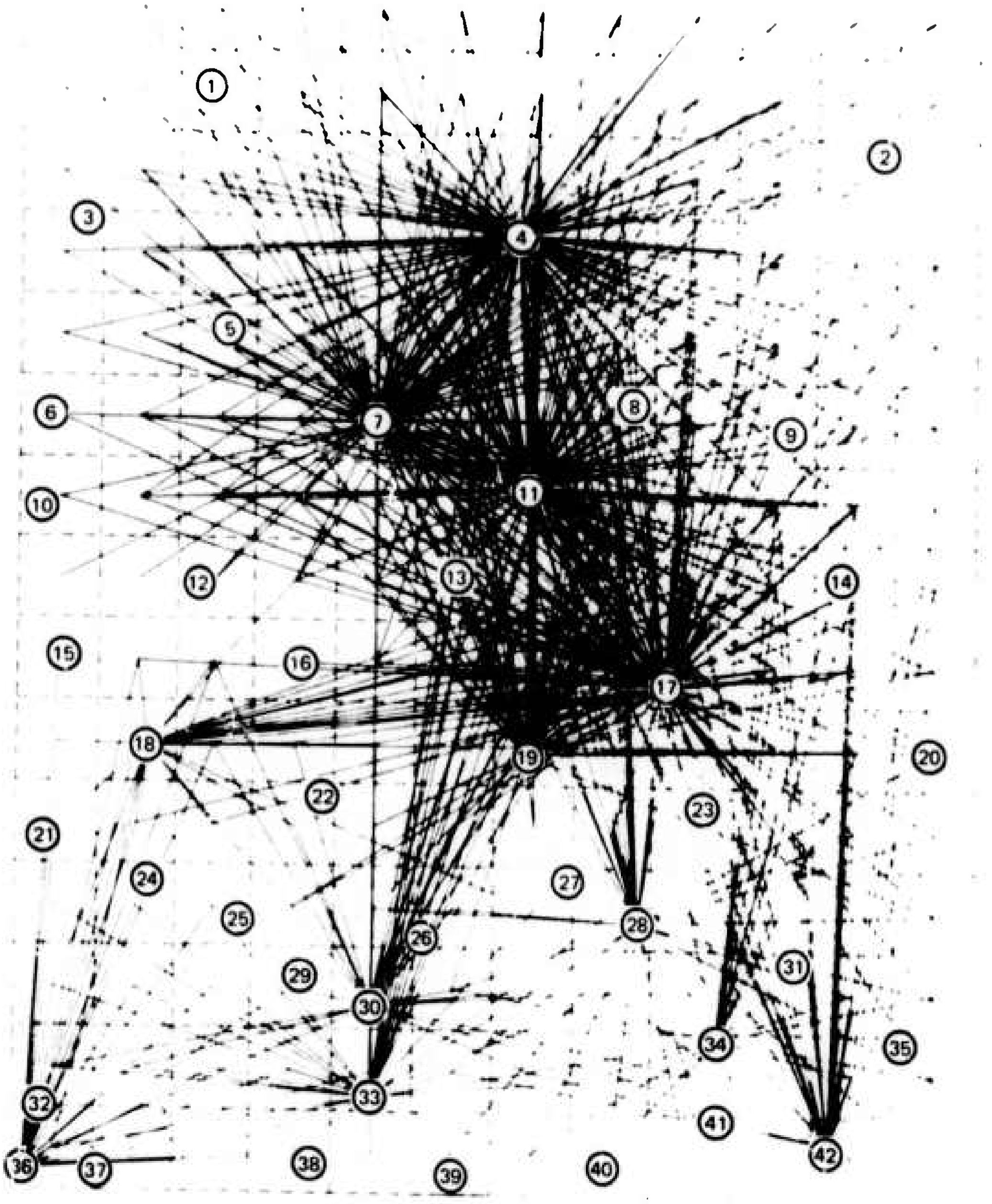


Figure 11.3: Repeater Covering for Example

STEP III

In this step, the stations transmit the structure information to repeaters and test each path in both directions.

Chapter 12
PACKET RADIO SYSTEM CHANNEL CONFIGURATION

12.1 Split versus Common Channel

Apart from the suitability for mobile terminals, random access schemes offer an attractive alternative to fixed assignment of channel capacity (FDM, TDMA) for applications characterized by traffic of a bursty nature. (That is, when the traffic requirements of users can be characterized as having a high peak to average data rate.) This is because at any given time, the capacity assigned to non active users is not utilized, whereas the active users experience relatively long delays due to the low data rate available to each.

We pursue this same argument one step further and investigate for the packet radio system whether we should have two channels, one for transmission from terminals to stations and the second in the reverse direction; or alternatively whether we should dynamically share the total capacity (common channel). This problem was investigated for a single hop network in which n stations communicate with an infinite number of terminals using the slotted ALOHA random access scheme [28]. In the model it is assumed that all stations and terminals are within an effective transmission range of each other, that the processes of packet originations and packet originations plus retransmissions are Poisson, and that there is a ratio a of the rate of packets which originates from stations to the rate which originates from terminals.

Figure 12.1 shows the comparison of the maximum effective utilization of the two configurations as a function of a with the number of stations, n, as a parameter. The subscripts s and c denote the split (into equal parts) and common configurations, respectively. The conclusion from this is that if the ratio a is not known or if it varies, it is preferable to share dynamically the total capacity. Figure 12.2 shows an example of the average delay of a packet in the system (weighted average of packet in the two directions) as a function of the total throughput, for the case $a = 10$. The difference in the packet transmission time (slot) due to the difference in the data rates of the two configurations has been taken into account. The superiority of the common channel configuration in this case ($a = 10$) is clearly demonstrated.

12.2 Directional Antennas and Multiple Transmitters

Another problem related to channel configuration is the possible use of directional antennas by repeaters and/or stations and the advantage (if any) of using multiple directional

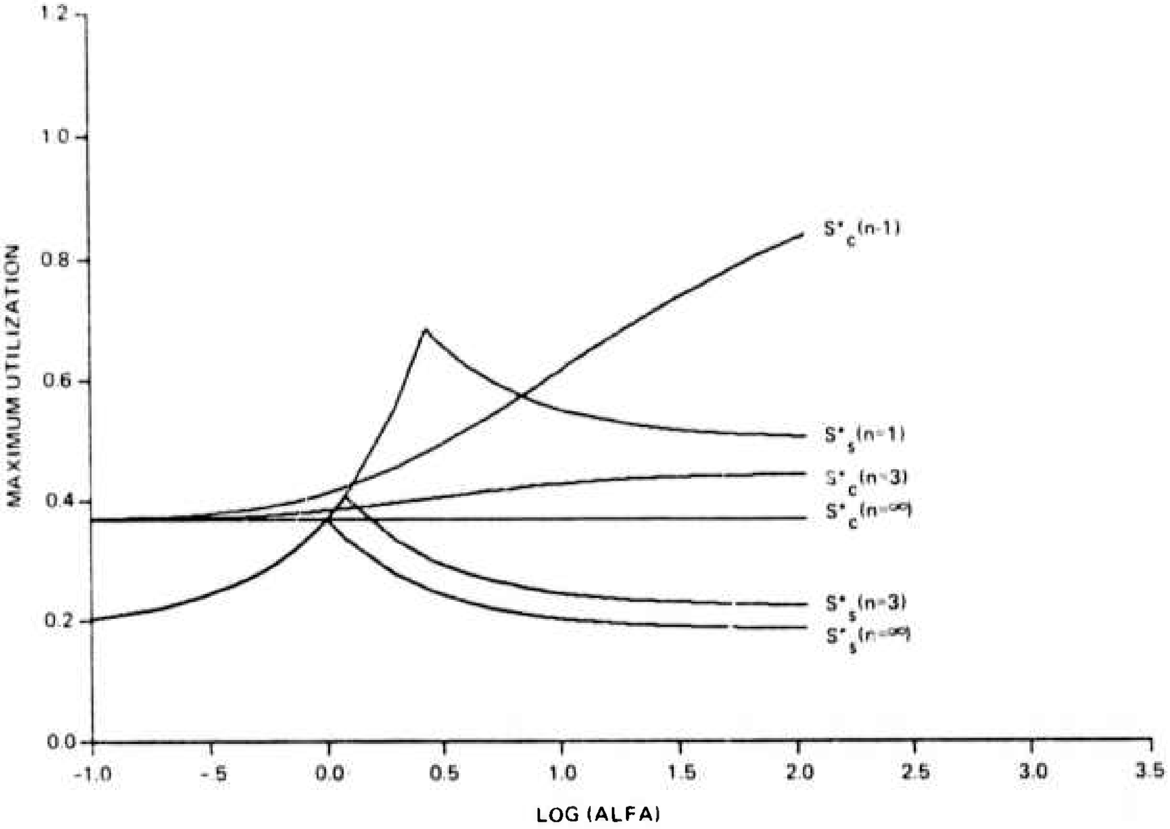


Figure 12.1: Maximum Utilization vs. Split and Combined Channel Parameters

transmitters. This problem was investigated for a 2-hop and single station packet-radio network [29]. The investigation was done assuming separate channels from station to terminals and from terminals to station, and for the slotted ALOHA random access scheme.

12.2.1 Transmission From Terminals To Station

Consider a 2-hop system with m repeaters and a single station as shown in Figure 12.3. The traffic originates from terminals and is destined to the station. A terminal transmits its packets to a repeater (hop 1), which in turn transmits the packets to the station (hop 2). The transmission protocol is as follows: when a packet becomes ready for transmission, it is transmitted into the next slot; the device then times out waiting for an ack, and if one is not received the packet is retransmitted at a future random slot.

We use the following assumptions. The combined process of packet originations and packet retransmissions, from each set of terminals to a repeater, is Poisson. The probabilities of transmission by a repeater into different slots are independent. The probability of transmission by two or more repeaters into a randomly chosen slot are mutually independent; and the probability of transmission into a random slot by a terminal and by a repeater are

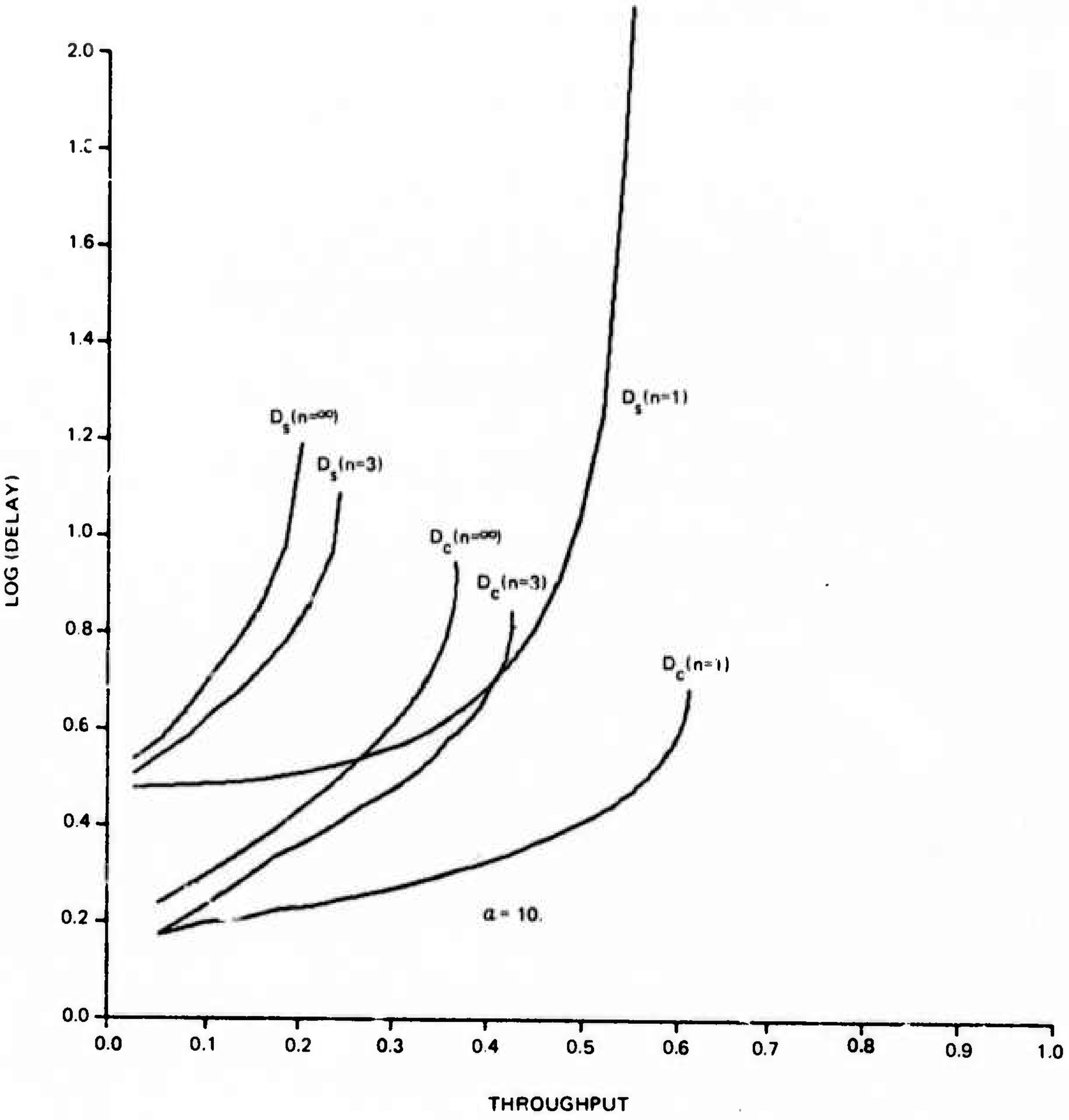


Figure 12.2: Delay vs. Throughput, a = 2.5

independent. Furthermore, we assume that the terminal transmission range is short, so that it can reach only one repeater. On the other hand, the transmission from a repeater to the station can interface with the transmission of terminals to l-1 other repeaters; 1 ⩽ 1 ⩽ m.

The effect of directional antennas at repeaters is that the transmission from repeaters to the station is directed towards the station and does not interfere with the transmission of terminals to other repeaters. Thus, it is the special case with l = 1. We notice, however, that directional antennas do not increase the capacity of the hop from repeaters to station

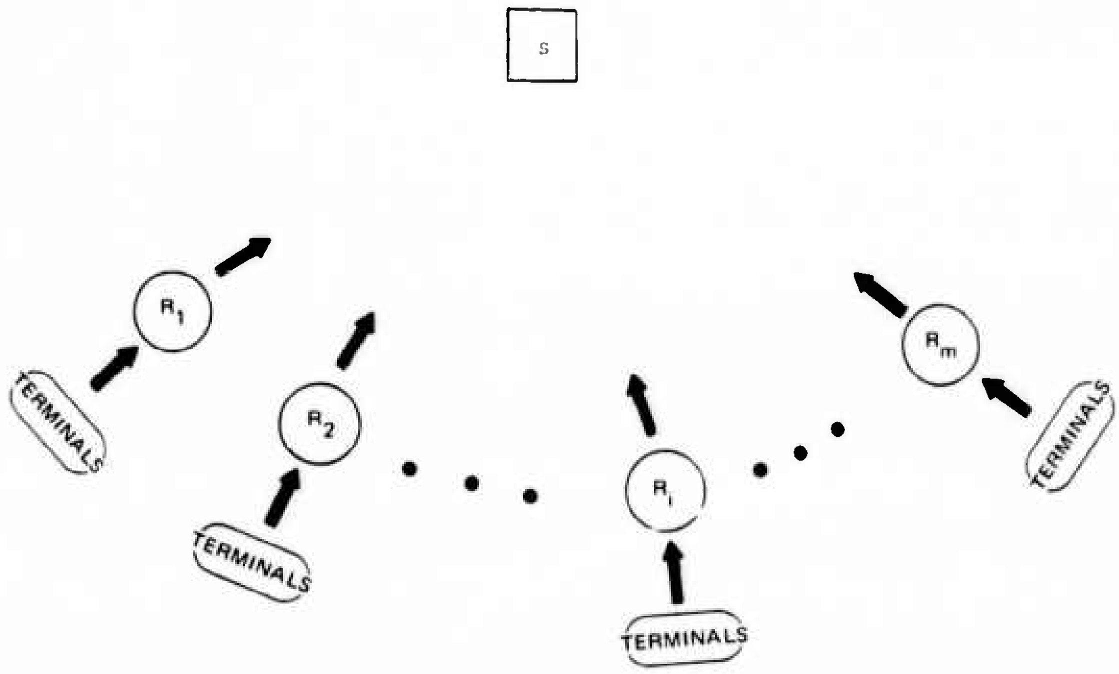


Figure 12.3: Transmission from Terminals to Station

because all antennas are directed towards the same physical location where the station is placed and where the conflicts may occur.

Figure 12.4 shows the capacity of the system as a function of the number of repeat:rs, m, for I = m and I = 1, which is equivalent to omnidirectional and directional antennas respectively. One can see that there is a significant gain in capacity when using directional antennas only when m = 2, and a small gain for m = 3; for m ⩾ 4 the capacity of the system does not increase.

As far as the number of repeaters is concerned, one can see 2 or 3 repeaters would be a good design; and additional repeaters that may be added because of other considerations (such as area coverage) will result in a reduction in the system capacity. Another problem investigated is the critical hop. That is, when the capacity of the system is reached, it is due to the saturation of the hop from terminals to repeaters or that from repeaters to the station. The results demonstrate that when the number of repeaters, m, is small the critical hop is from terminals to repeaters, whereas when m is large the critical hop is from repeaters to station. The exact number at which the change occurs depends on the interference parameter I.

12.2.2 Transmission From Station To Terminals

In this section, we consider the second channel which is used for transmission from the station to terminals via repeaters. It is assumed that the effective transmission range of

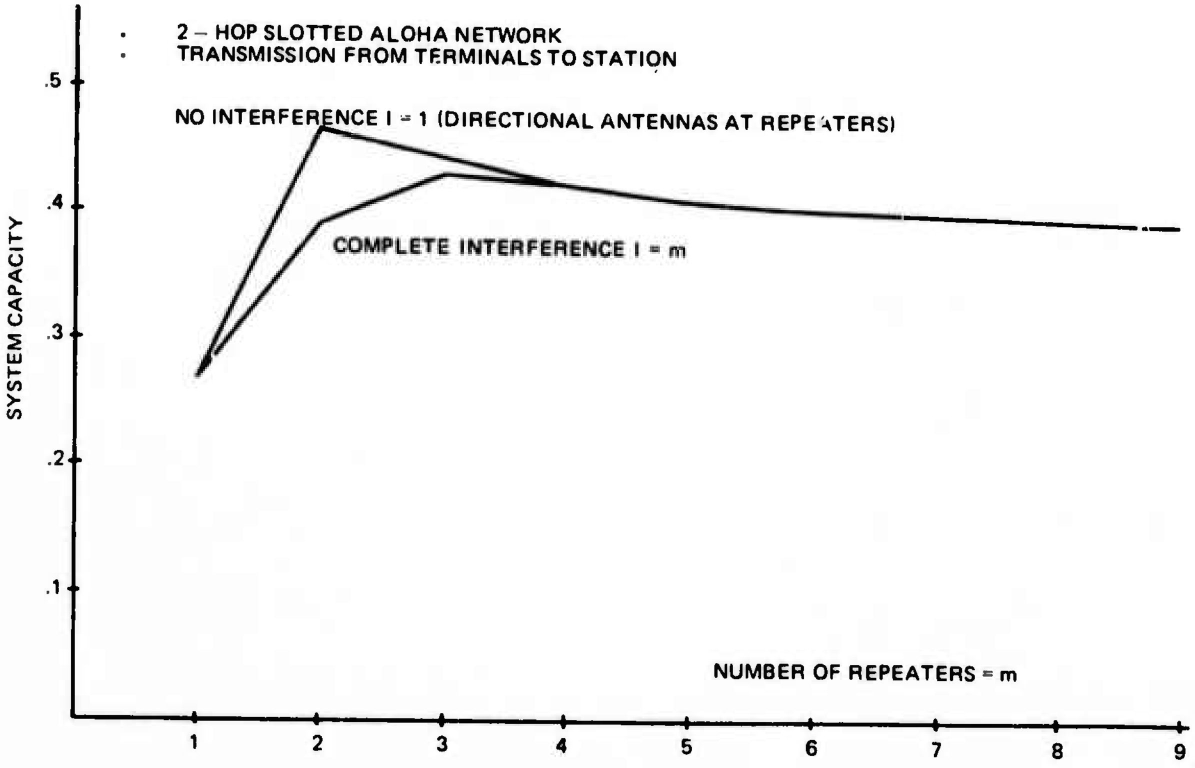


Figure 12.4: Network Throughput vs. Number of Repeaters: Directional and Non-Directional Antennas

the station is such that it interferes with the transmission from repeaters to terminals. However, we assume that terminals are not designed to directly receive from the station. We use the same assumptions as in the previous section. A transmission from the station to R_i can be interfered with by transmissions from the I repeaters in the interfering set of R_i when these repeaters transmit to their terminals (T's). A transmission from R_i to T can be interfered by a transmission from the station to any repeater or by the I – 1, excluding R_i, repeaters in the interfering set of R_i. For consistency with the interference model in the previous section we assume the same energy-per-bit-to-noise-density for detection with equal error rates, by the repeaters and by the terminal and that the repeater uses a higher transmitter power than terminals.

Figure 12.5 shows the capacity of the system as a function of m for I = 1 and I = m, both for an omnidirectional and directional antenna from the station to repeaters. Further investigations for this case were performed and the conclusions follow.

a. The interference of the station with the transmission of repeaters to terminals significantly reduces the system capacity. Thus, if possible, it is important to enable terminals to receive such transmissions directly, without retransmission by the repeater.

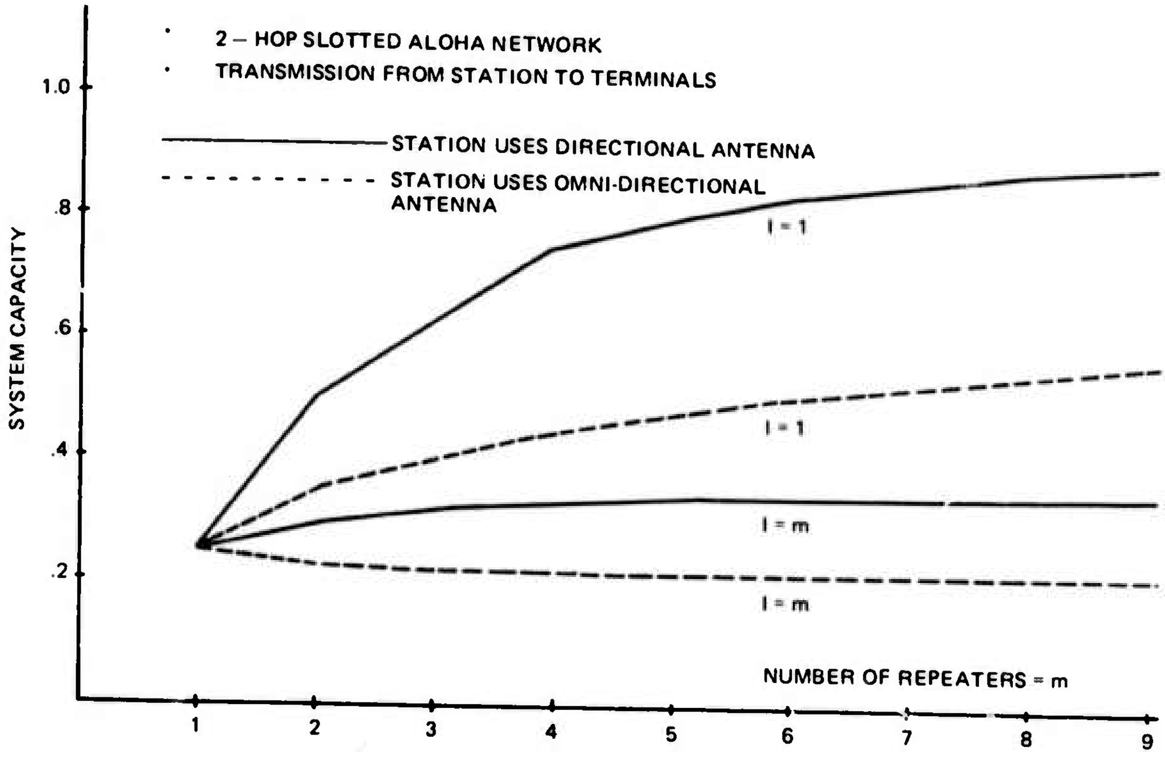


Figure 12.5: Number of Repeaters vs. Capacity

b. The system capacity is reduced substantially when the interference level between repeaters is increased. Note that this is not the case when transmitting to the station. Consequently, it is important to reduce the interference factor by a mechanism such as adaptive power.

c. A directional antenna at the station significantly increases the system capacity when the interference level between repeaters is low to moderate. This is not the case when the interference level is high, since the throughput on the hop from repeaters to terminals is limited due to this interference.

d. When the station has directional antennas, then multiple transmitters and antennas may further increase significantly system capacity. In this case one can obtain a capacity greater than 1.

Chapter 13
PACKET RADIO SYSTEM ROUTING AND ACKNOWLEDGEMENT CONSIDERATIONS

13.1 Routing Problems

Problems that arise in controlling traffic in a broadcast net include:

a. A packet transmitted can be received by many repeaters or stations or not be received by any.

b. Many copies of the same packet can circulate in the broadcast network.

c. Many copies of the same packet can enter the point-to-point network at different stations.

Indications of the consequences of not imposing a suitable flow control mechanism can be observed from idealized combinatorial models.

In these ideal models, the repeaters are located at corner points of an infinite square grid and time is broken into unit intervals, each slotted into segments. A packet transmitted by a repeater can be received only by its four nearest neighbors. If a packet is correctly received by a repeater, it is retransmitted within the next unit interval of time at a random time slot within the interval. Suppose now that *a single packet* originates at the origin and that the transmission plus the propagation time falls within one unit interval of time. Then after n intervals of time:

a. The number of repeaters which receive the packet for the first time, B(n), is:

$$B(n) = 4n,\ n > 1 \quad B(0) = 1$$

b. The number of repeaters through which the packet passed, A(n), is:

$$A(n) = \sum_{j=0}^{n} B(j) = 2n^2 + 2n + 1,\ n \geq 0$$

c. if we assume that a repeater can receive and relay a large number of packets within the same time interval, the number of copies of the same packet received by a repeater at coordinates (d,j) after d + 2k units of time is:

$$N_j^d\ (d =+ 2k) = \binom{d+2k}{k+j}\binom{d+2k}{k} \xrightarrow{\text{for large k}} 2^{4k}$$

where d is the number of units of time that the packet requires to arrive from the origin to the repeater, and j is the horizontal number of units.

Unless adequate steps are taken, the explosive proliferation of redundant packets will severely limit the capacity of the system. One can now recognize two somewhat distinct routing and control problems:

a. To ensure that a packet originating from a terminal arrives at a station, preferably using the most efficient (shortest) path,

b. To suppress copies of the same packet from being indefinitely repeated in the network, either by being propagated in endless cycles of repeaters or by being propagated for a very long distance.

13.2 Proposed Routing Techniques

There are two key objectives in developing a routing procedure for the packet radio system. First, we must assure, with high probability, that a message launched into the net from an arbitrary point will reach its destination. Second, we must guarantee that *a large number* of messages will be able to be transmitted through the network with a relatively small time delay. The first goal may be thought of as a *connectivity* or *reliability* issue, while the second is an *efficiency* consideration.

13.2.1 Undirected Routing

A rudimentary, but workable, routing technique to achieve connectivity at low traffic levels can be simply constructed by using a maximum handover number [4] and saving unique identifiers of packets at each repeater for specified periods of time. The handover number is used to guarantee that any packet cannot be indefinitely propagated in the net. Each time a packet is transmitted in the net, a handover number in the header is incremented by one. When the handover number reaches an assigned maximum, the packet is no longer repeated and that copy of the packet is dropped from the net. Thus, the packet is "aged" each time it is repeated until it reaches its destination or is dropped because of excessive age.

If the maximum handover number is set large, extensive artificial traffic may be generated in areas where there is a high density of repeaters. On the other hand, if it is set small, packets from remote areas may never arrive at stations. This problem can be resolved as follows: We assume that every repeater can calculate its approximate distance in numbers of hops to stations by observing response packets. The first repeater which received the packet from a terminal sets the maximum handover number based on its calculated distance from the station. The number is then decremented by one each time it is relayed through any other repeater. The packet is dropped when the number reduces to zero. When a station transmits a packet, it will set the maximum handover number by "knowing" the approximate radius in "repeaters" in its region.

Even if a packet is dropped after a large number of transmissions, local controls are needed to prevent packets from being successively "bounced" between two or a small number of repeaters which repeat everything they correctly receive. (Such a phenomena is called "cycling" or "looping.") A simple mechanism to prevent this occurrence is for repeaters to store for a fixed period of time entire packets, headers, or even a field within the header that uniquely identifies a packet. A repeater would then compare the identifier of any received packet against the identifiers in storage at the repeater. If a match occurred, the associated packet would not be repeated.

The time allotted for storage of any packet identifier would depend on the amount of available storage at a repeater and the number of bits required to uniquely identify the packet. For example, more than 4K packets could be uniquely identified with 12 bit words. Thus, 4K of storage could contain identifiers for more than 300 packets. With a 500 Kbps repeater to repeater common channel for broadcast and receive and 1,000 bit packets, this would be sufficient storage for over 1.5 seconds of transmission if the channel were used at full rate. Assuming a single hop would require about 20 milliseconds of transmission and retransmission time, a maximum hop number of 20 would guarantee that any packet would be dropped from the system because of an excessive number of retransmissions long before it could return to a previously used repeater not containing the packet identifier.

The combination of loop prevention and packet aging with otherwise indiscriminate repetition of packets by repeaters will enable a packet to travel, on every available path, a maximum distance away from its origin equal to its original handover number. Thus, if the maximum handover number is larger than the minimum number of hops between the terminal and the nearest station, a packet accepted into the net should reach its destination. Unfortunately, with this scheme, copies of the packet will also reach many other points, with each repetition occupying valuable channel capacity. However, if those packets for which adequate capacity is not available are prevented from entering the net, the network will appear highly reliable to accepted packets.

The above routing scheme is an *undirected*, completely distributed procedure. Each repeater is in total control of packets sent to it, and the stations play no active part in the system's routing decisions. (They must still play a role in flow control.) In the above procedure, no advantage is taken of the fact that most traffic is destined for a station, either as a terminus or as an intermediate point for communication with elements of a different network. Also, the superior speed and memory space of the station is ignored. For efficiency, one is therefore led to investigate *directed* (hierarchical) routing procedures.

13.2.2 Directed Routing

A directed routing procedure utilizes the stations to periodically structure the network for efficient flow paths. Stations periodically transmit routing packets called *labels* to repeaters to form, functionally, a hierarchical point-to-point network as shown in Figure 13.1. Each label includes the following information:

a. A specific address of the repeater for routing purposes

b. The minimum number of hops to the nearest station

c. The specific addresses of *all* repeaters on a shortest path to the station (In particular, the label contains the address of the repeater to which a packet should preferably be transmitted when destined to the station.)

When relaying a packet to its destination, the repeater addresses the packet to the next repeater along the preferred path. Only this addressed repeater will repeat the packet and only when this mechanism fails will other repeaters relay the message.

For simplicity, we describe routing for the case of a single station network. A label of repeater R_i of hierarchy level j will be denoted by L_{ij}; i,j>1. The station will have the label L_{11}. L°_{ij} will denote the label of the repeater which is the "nearest available" to the communicating terminal.

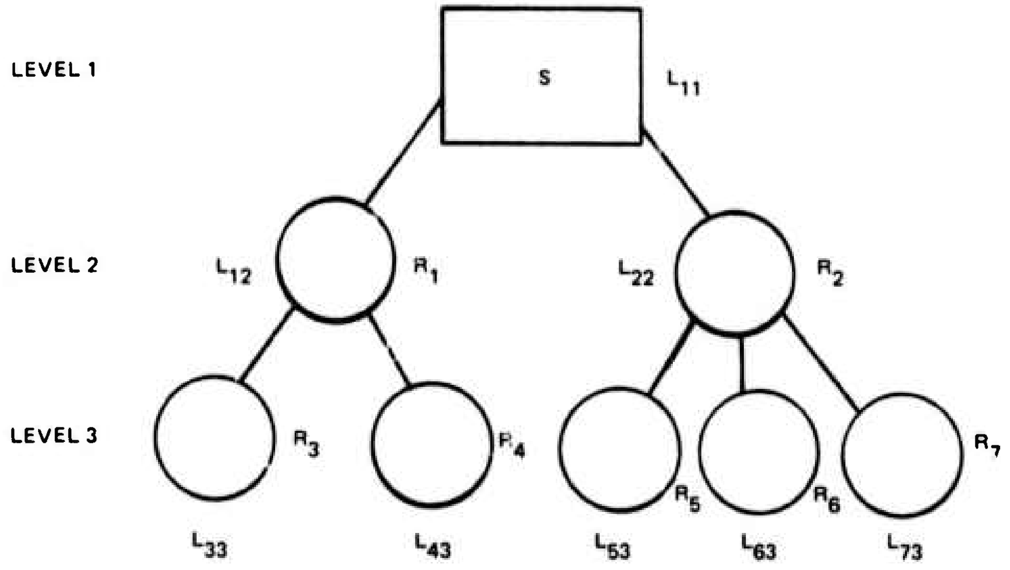


Figure 13.1: Hierarchical Labeling For Directed Routing Algorithms

A label is composed of H subfields, where H is the maximum number of hierarchy levels (H-1 is the maximum number of hops on the shortest path between any repeater and the station). Every subfield has three possible entries, blank (BLK), a serial number (SER), or ALL. L_{ij} has j entries of SER's and (H-j) BLK's as shown in Figure 13.2.

1	2		j-1	j	j+1		H
SER	SER	. . . .	SER	SER	BLK	. . . .	BLK
j serial numbers					(H-j) blanks		

Figure 13.2: Definition of Packet Label

We say that L_{ij} "homes" on L_{kp}, $h(L_{ij}) = L_{kp}$, if p = j-1 and the first j-1 subfields of both are identical. If two repeaters at level j home on the same repeater, their label will differ only in the entry to subfield j.

As an example, if we use 3 bits per subfield, the labels of the station and the repeaters of the network shown in Figure 13.1 are as follows:

	Subfield 1	**Subfield 2**	**Subfield 3**
L_{11}	0 0 1	0 0 0	0 0 0
L_{12}	0 0 1	0 0 1	0 0 0
L_{22}	0 0 1	0 1 0	0 0 0
L_{33}	0 0 1	0 0 1	0 0 1
L_{43}	0 0 1	0 0 1	0 1 0
L_{53}	0 0 1	0 1 0	0 0 1
L_{63}	0 0 1	0 1 0	0 1 0
L_{73}	0 0 1	0 1 0	0 1 1

In this example, a subfield in which all bits are "0" is considered "blank." Note that all entries in Subfield 1 are the same since all repeaters home (eventually) on the same station.

The packet header, shown in Figure 13.3, includes the following information.

L_{kn}	L°_{ij}	OTHER HEADERS AND PACKET INFORMATION
TO	LABEL OF NEAREST REPEATER TO THE TERMINAL	

Figure 13.3: Routing Information Contained in Packet Header

L_{kn} is the label of the repeater to which the packet is currently addressed. The complete packet will *always* be transmitted to a *specific device;* other devices which may receive the packet will drop it. The shortest path from a terminal to the station consists of L°_{ij}, $h(L^{\circ}_{ij})$, $h(h(L^{\circ}_{ij}))$, up to L_{11}, in the given order, and in the reverse order when routing from station to terminal. When a specific repeater along the shortest path is not known (by the terminal) or not available, then the terminal or repeater (which has the packet) will transmit *only the header part* of the packet, trying to identify a specific repeater. In that case, the label L_{kn} will include some entries ALL. To see how the proposed routing technique would operate, we trace the sequence of steps performed when a terminal attempts to transmit a packet to the station.

When a previously silent terminal begins to communicate, it first identifies a repeater or a station in its area. It transmits only the header part of the packet with all entries in L_{kn} set to ALL. The header is addressed to all repeaters and stations that can hear the terminal. A device which correctly receives this header substitutes its label in the space L_{kn} and repeats the header. This particular L_{kn} is also L°_{kn} and will be used by the terminal to transmit all packets during this period of communication. If a terminal is stationary, it can store this label for future transmissions. L°_{kn} begins to transmit the complete packet along the shortest path to the station.

Suppose that L_{ij} along the shortest path is not successful in transmitting the packet to $h(L_{ij})$. Then L_{ij} begins the search stage of trying to identify another repeater. In the first step, it tries to identify a repeater which is in level $p \leq j-1$. This is done by using the label shown in Figure 13.4.

1	2	3		j-1	j	j+1		
SER	ALL	ALL	. . .	ALL	BLK	BLK	. . .	BLK

Figure 13.4: Label Used In Search Process

The header is addressed to all repeaters in levels 2 to j-1, which eventually home on L_{11}. If this step is not successful, in the second (last) step, L_{ij} tries to identify any available repeater by using a label in which the first entry is SER and all other entries are ALL. When a specific repeater is identified and receives the packet, it transmits the packet on the shortest path from its location.

Note that if repeaters have sufficient storage, they can save alternative labels and thus reduce the necessity of searching for a specific repeater. Alternative solutions in which repeaters have multiple labels are also possible.

13.3 Acknowledgement Considerations

Acknowledgement procedures are necessary both as a guarantee that packets are not lost within the net and as a flow control mechanism to prevent retransmissions of packets from entering the net. Two types of acknowledgements are common in packet oriented systems:

a. Hop-by-Hop Acknowledgements (HBH Acks) are transmitted whenever a packet is received successfully by the next node on the transmission path.

b. End-to-End Acknowledgements (ETE Acks) are transmitted whenever a packet correctly reaches its final destination within the network.

In a point-to-point oriented network such as the ARPANET, HBH Acks are used to transfer responsibility (and thus open buffer space) for the packet from the transmitting node to the receiving node. This Ack insures prompt retransmission should parity errors or relay IMP buffer congestion occur. The ETE Ack serves as a flow regulator between source and destination and as a signal to the sending node that the final destination node has correctly received the message. Thus, the message may be dropped from storage at its origin.

Both types of Ack's serve to ensure message integrity and reliability. If there is a high probability of error free transmission per hop and the nodes have sufficient storage, the Hop-by-Hop scheme is not needed for the above purpose. Without an HBH Ack scheme, one would transmit the packet from its origin after a time out period expired. One introduced the HBH Ack to decrease the delay caused by retransmissions at the expense of added overhead for acknowledgements. In the ARPANET, this added overhead is kept small by "piggybacking" acknowledgements whenever possible on information packets flowing in the reverse direction. In the packet radio system, the overhead can be kept small by listening, whenever possible, for the next repetition of the packet on the common channel instead of generating a sepa.ate acknowledgement packet.

The value of an End-to-End acknowledgement is sufficiently great that it can be assumed present *a priori*. However, the additional use of a Hop-by-Hop acknowledgement is not as clear. Therefore, in this section, we examine the question of whether the ETE Ack is sufficient, or whether one needs a Hop-by-Hop (HBH) acknowledgement in addition. The problem is therefore whether an HBH Ack is superior to an ETE Ack with respect to throughput and delay, since the ETE Ack ensures message integrity. It is noted that the routing and flow control by devices in the network depend on the type of acknowledgement scheme used.

We consider a simple case where (n-1) repeaters separate the packet radio terminal from the destination station. Assuming that the terminal is at a distance of "one hop" from the first repeater, one obtains the n-hop system shown in Figure 13.5.

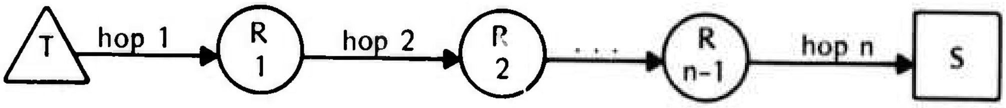


Figure 13.5: Multi-Hop Stream

A simple model is used to evaluate the total average delay that a packet encounters in the n-hop system when using HBH and ETE acknowledgement schemes. When the ETE acknowledgement scheme is used, every repeater transmits the packet a single time. If the packet does not reach the station, retransmission is originated by the terminal. The ETE acknowledgement is sent from the station. In the HBH scheme, repeaters store and retransmit the packet until positively acknowledged from the next repeater stage.

If, after a terminal (or a repeater in the HBH case) transmits the packet, an acknowledgement does not arrive within a specified period of time, it retransmits the packet. The waiting period is composed of the time for the acknowledgement to arrive when no conflicts occur plus a random time for avoiding repeater conflicts.

Two different schemes for ETE acknowledgement and one scheme for HBH acknowledgement have been studied. Curves for the total average delay as a function of the number of hops and the probability of successful transmission per hop are obtained. Two cases are considered: One in which the probability of success is constant along the path and another in which the probability of success decreases linearly as the packet approaches the station. Finally, channel utilizations are compared when using ALOHA [1] random access modes of operation.

It has been demonstrated that the HBH scheme is superior in terms of delay or channel utilization. This conclusion becomes significant when the number of hops increases or when the probability of successful transmission is low. For example, in a five hop system, if the probability of success per hop is 0.7, then the total average delay is 12.5 and 53 packet transmission times for the HBH and ETE acknowledgement schemes, respectively.

Chapter 14
RANGE, POWER, DATA RATE AND CAPACITY CONSIDERATIONS

14.1 Transmission Range and Network Interference

A variety of situations is possible concerning the range and interference patterns of devices. For example, with identical r.f. elements and similar antenna placements, Repeater-to-Repeater range is the same as Terminal-to-Repeater range. This, however, is not always a necessary limitation since repeaters can be placed on elevated areas and can have more power than terminals (especially hand held terminals). Thus, if repeaters are allocated for *area coverage of terminals*, the range will be higher than terminal range and higher network connectivity or device interference will result.

The problem which then arises is to determine the impact of this interference on system performance. Alternatively, one may seek to reduce repeater transmission power when transmitting in the repeater-station network. As an indication of the tradeoffs that occur, common channels and the single date rates (CCSDR) were simulated, one with High Interference CCSDR (HI), and the other with Low Interference CCSDR (LI). As a first step, the routing labels of the two systems were the same and are shown in Figure 14.2. The interference of the CCSDR (LI) system is shown in Figure 14.1 and the interference of the CCSDR (HI) system in Figure 14.3. (Figure 14.3 shows only the connectivity for two devices in the network.) A different label assignment for the high interference system is shown in Figure 14.4.

The results are shown in Figure 14.5 and Table 14.1. Figure 14.5 shows the throughput of the two systems as a function of time while Table 14.1 summarizes other measures of performance. The third row of Table 14.1 summarizes performance of the high interference system under an improved set of repeater labels. It is clear that the high interference system has better performance than the low interference system. The only measure of the low interference system which is better is terminal blocking which is a direct result of the low interference feature. In fact, CCSDR (LI) is saturated at the offered traffic rate. This can be seen from the fact that the throughput is decreasing as a function of time, the relatively high total loss, and the low station response.* The CCSDR (HI) with improved labels, compared in Table 14.1, has better performance than the other two systems. This indi-

*The average number of station response packets assumed for these studies is 2.0.

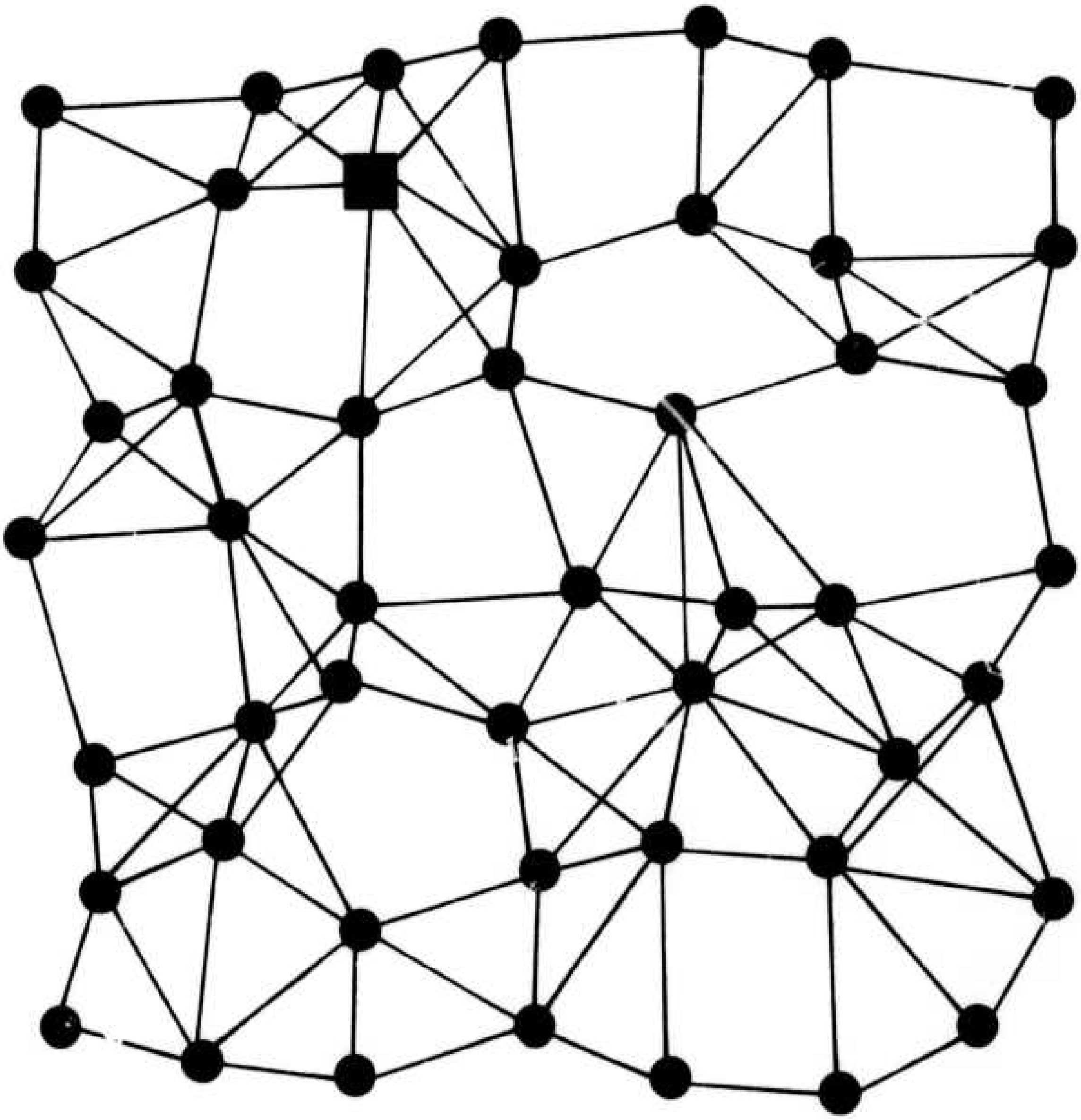


Figure 14.1: Connectivity of Repeaters and Stations

cates the importance of proper labeling. The experiments of this section demonstrate that it is preferable to use high transmitter power to obtain long repeater range, despite the network interference that results.

14.2 Single Versus Dual Data Signaling Rates Networks

The previous results demonstrate that better performance is obtained when repeaters and stations use high power to obtain long range despite the interference that results. We now examine the problem of whether repeaters and stations should use their fixed power budgets to obtain a long range with a low data rate channel or have a short range with a high data rate channel. The following systems were studied.

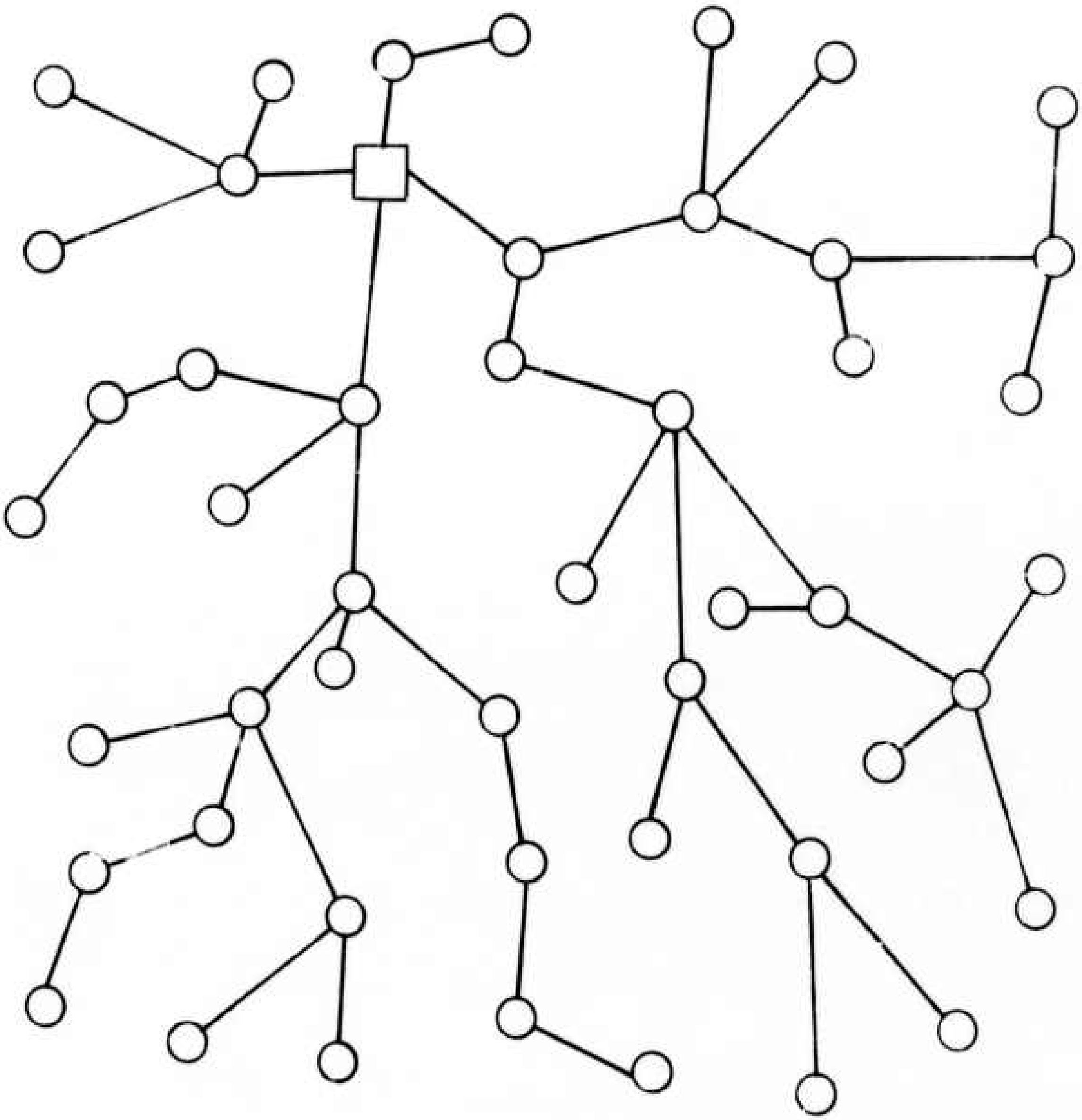


Figure 14.2: Hierarchical Labeling Scheme

- The CCSDR (HI) of the previous section with improved label. to take advantage of the high range to improve the routing labels of repeaters and obtain fewer hierarchy leveis which we denote by CCSDR. The routing labels used are shown in Figure 14.4, and the connectivity is shown in Figure 14.3.

- A Common Channel Two Data Rate (CCTDR) system with the routing labels as in Figure 14.2 and connectivity as in Figure 14.1.

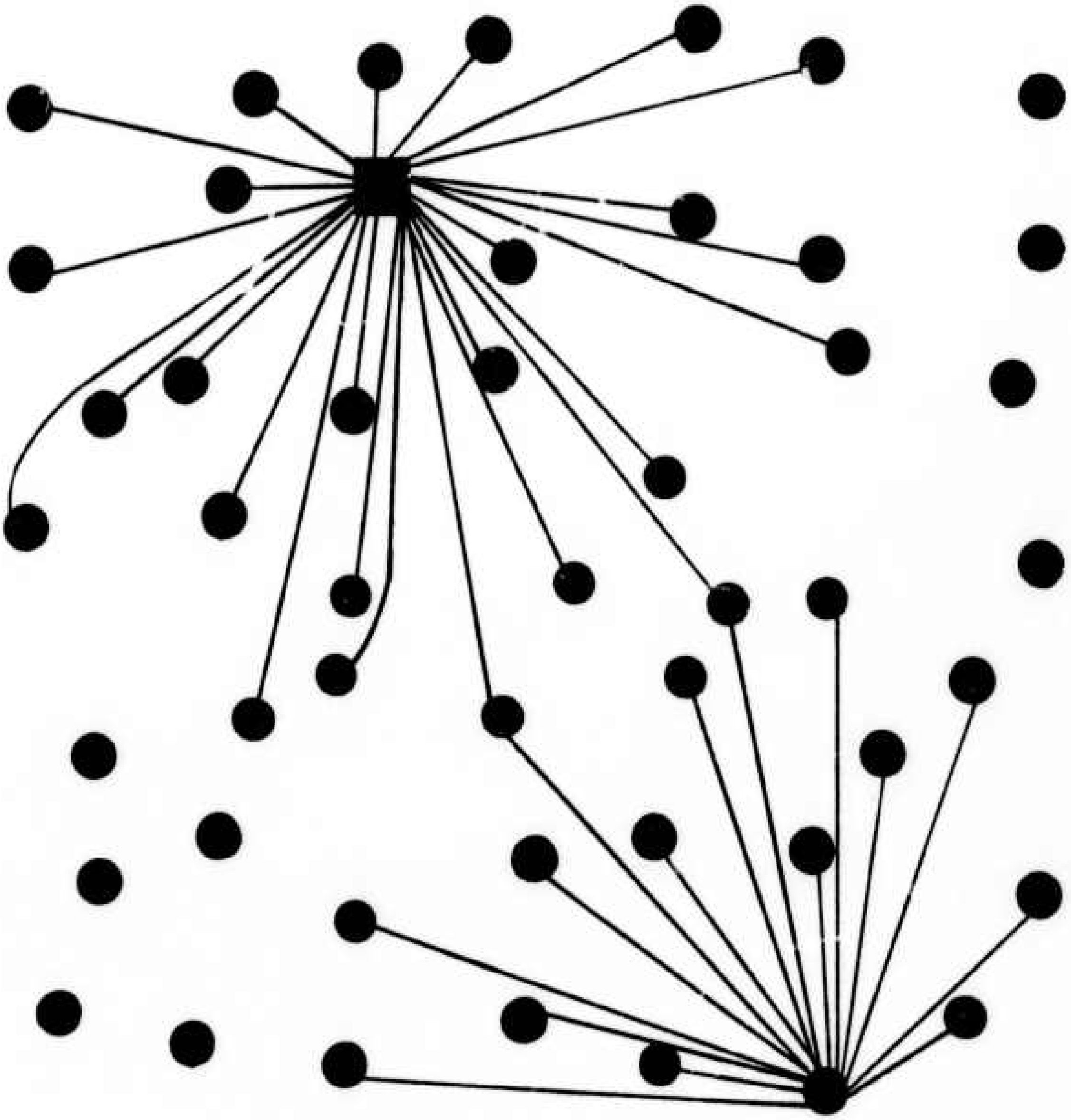


Figure 14.3: Interference of CCSDR (HI) System

In the CCTDR system, the terminal has a low data rate channel, the same rate as in the single data rate system, for communication with a repeater or station. Repeaters and stations have two data rates. The high data rate is used for communication in the repeater-station network. The two data rates use the same carrier frequency so that only one can be used at a time.

The two systems are tested with offered rates of 13% and 25%.* The throughput as a function of time for the two runs is shown in Figures 14.6 and 14.7, respectively; and

*In the simulation runs we used the inverse square law for the relation between data rate and distance, rather than the result in [9]; this, however, favors CCSDR.

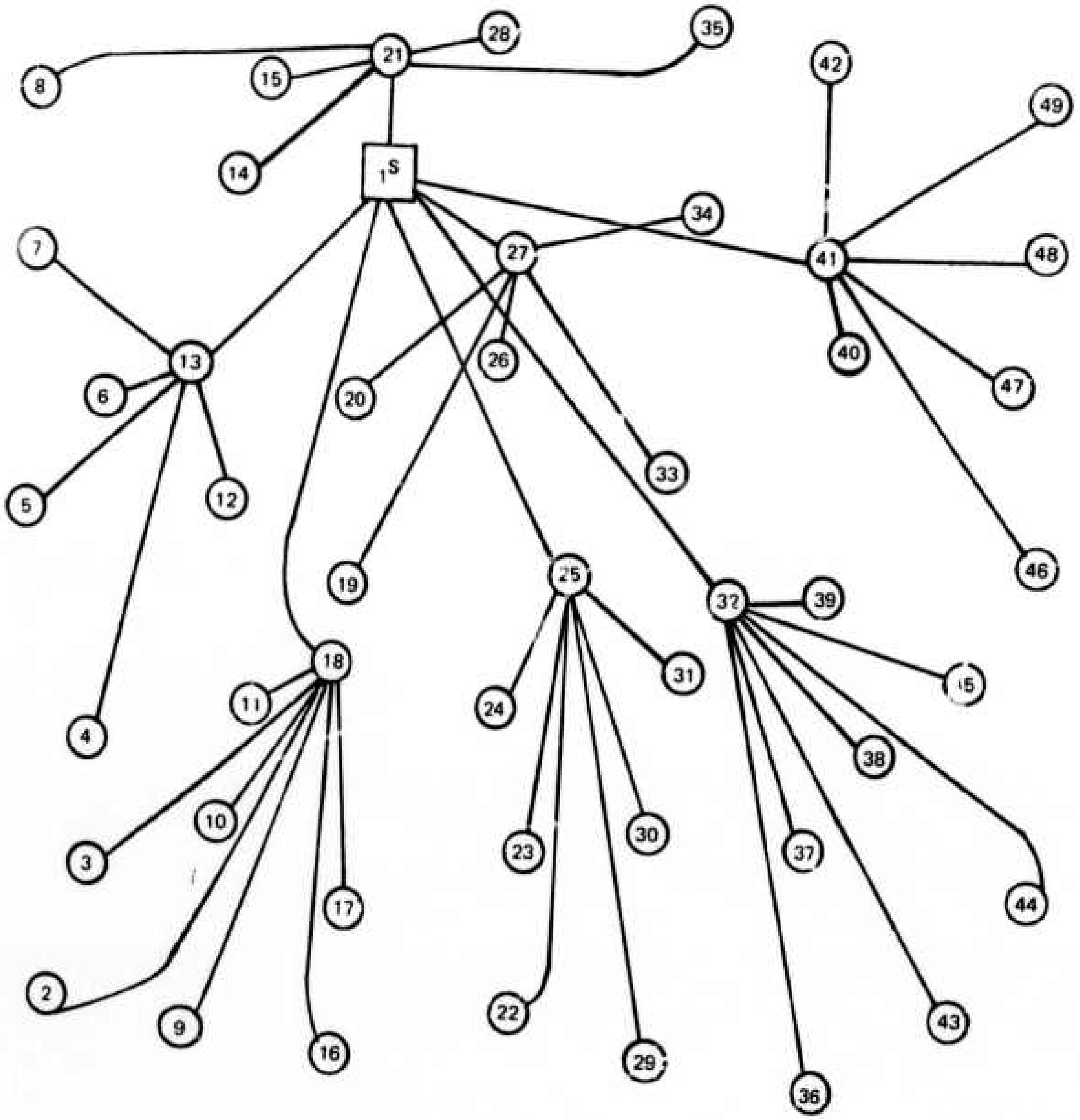


Figure 14.4: CCSDR (HI) System With Improved Labeling

the summary of other measures is given in Table 14.2. The comparison demonstrates that the CCTRD system is superior to the CCSDR system, in terms of throughput, delay, and other measures. One can see that the CCSDR system is saturated at an offered rate of about 13%.

14.2.1 Effect on Blocking Level

In Table 14.2, one can see that one reason for the relatively low throughput of the CCSDR system at an offered rate of 25% is due to blocking. Furthermore, the fraction of time

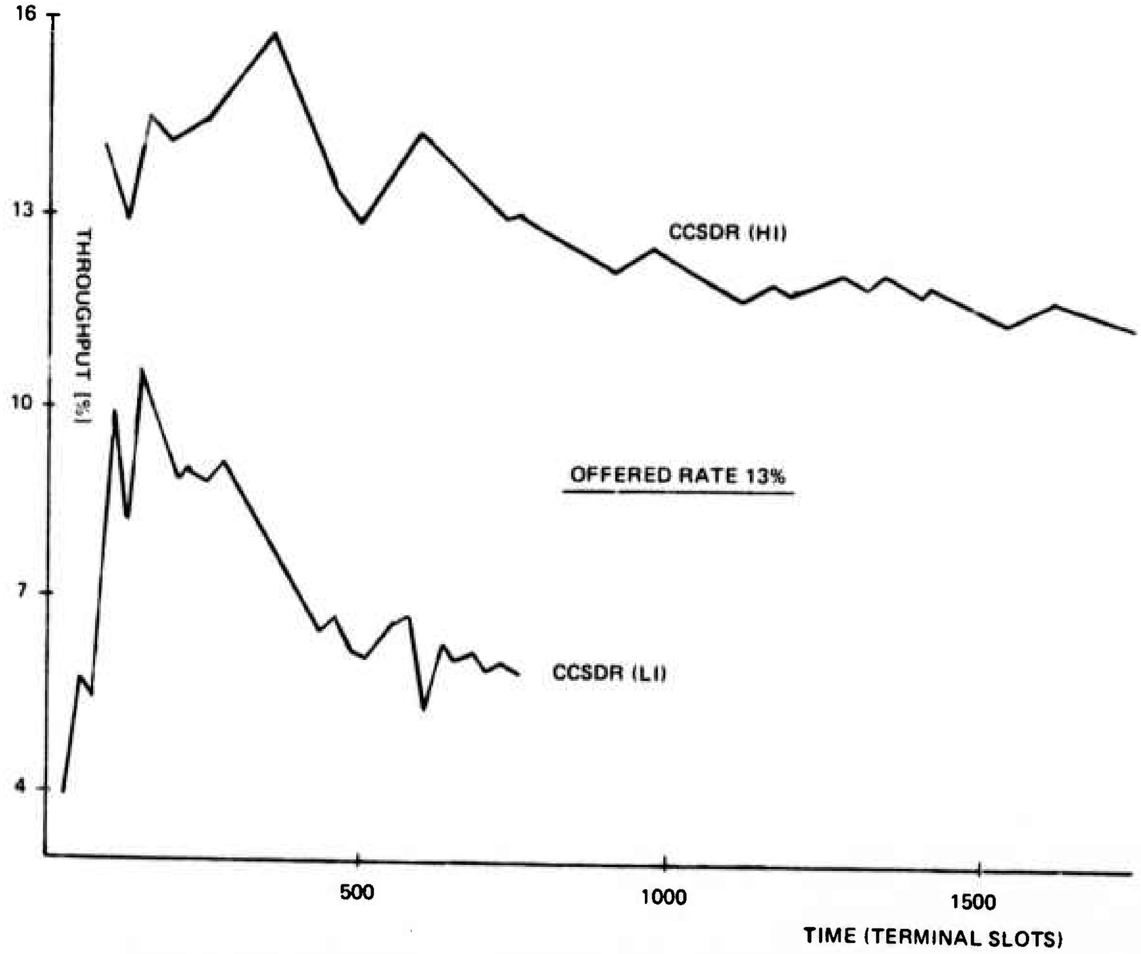


Figure 14.5: Throughput vs. Terminal Slots: CCSDR (HI) and CCSDR (LI)

Table 14.1: Effect of Range On Network Performance For Single Data Rate System

Interference Level	Offered Rate (%)	Throughput (%)	Delay Of IP To Station (Terminal Slots)	Rate Of Station Response	Prob. Station Busy	% Of IP Blocked	Total % Of IP Loss	Terminals Remaining
CCSDR (LI)	13	5.95	40.11	1.14	.53	2.98	32.83	13
CCSDR (HI)	13	10.55	23.93	1.81	.43	9.83	9.83	13
CCSDR (HI) (Improved Labels)	13	12.14	16.61	2.06	.50	10.63	11.41	10

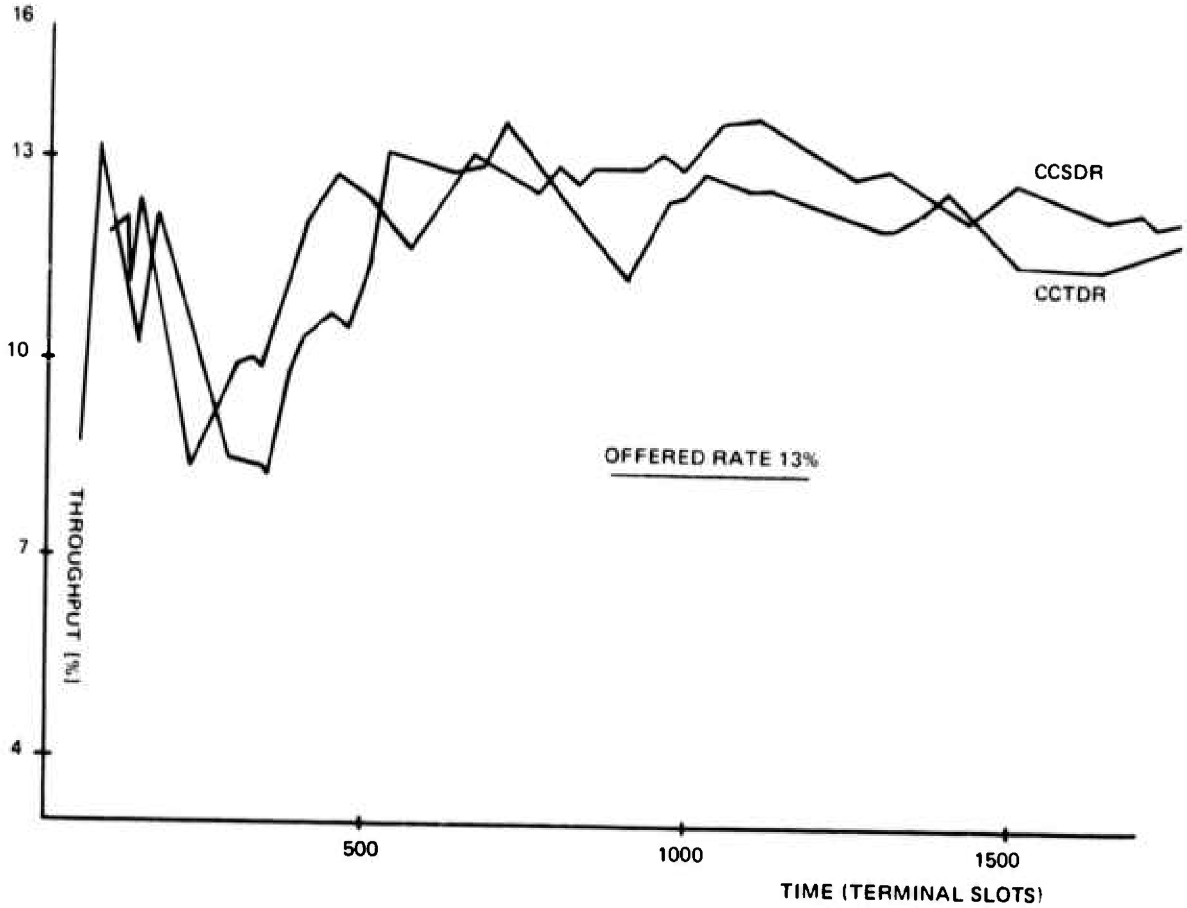


Figure 14.6: Throughput vs. Terminal Slots: 13% Rate

that the station is busy has decreased. This may suggest that the station may be able to handle more terminals providing they are able to enter the system. To examine this point, we ran the CCSDR system with an offered rate of 25%, and relaxed the constraint for entering the system. Rather than resulting in better performance, this step resulted in reduction in blocking and increase in delay. The throughput increased to 12.63%, the blocking decreased to 18.35% and the total loss decreased to 30.73%. On the other hand, the delay increased to 57.82, the fraction of time the station is busy increased to .57, and the rate of station response decreased to 1.32.

To conclude, when we enabled more terminals to enter the system, the throughput increased insignificantly, from 12.20% to 12.63%; on the other hand, the average packet delay *increased significantly*, from 34.97 to 57.82 terminal slots. This suggests that one of the important design problems in the packet radio system is the blocking level of terminals.

14.3 Throughput, Loss, and Delay of CCSDR and CCTDR Systems

Similar to curves of throughput versus channel traffic, for which the relation is known analytically [38], we can attempt to draw curves of system throughput vs. offered rate for

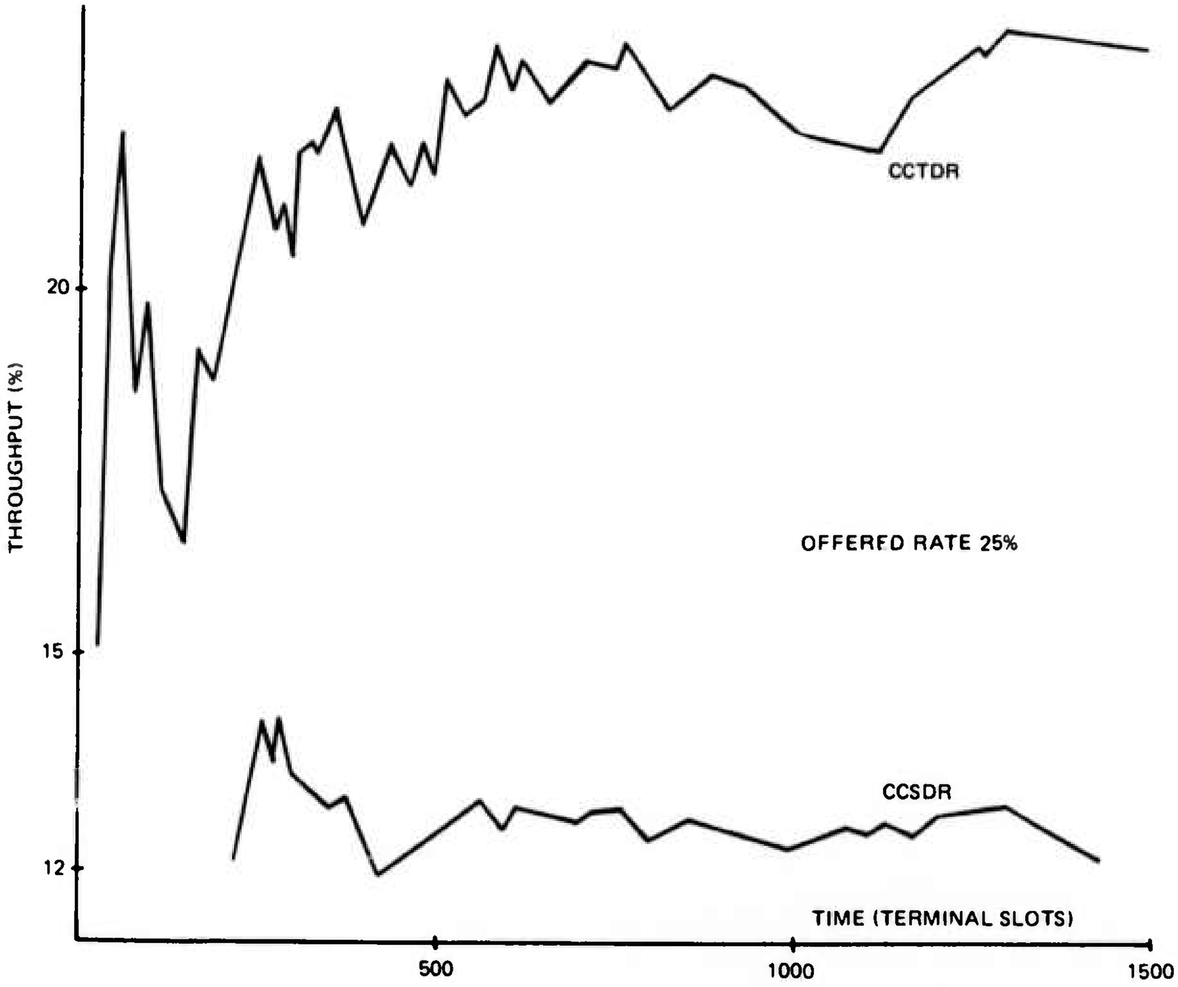


Figure 14.7: Throughput vs. Terminal Slots: 25% Rate

Table 14.2: Single Data Rate vs. Dual Data Rate System Performance

	Offered Rate (%)	Throughput (%)	Delay Of IP To Station (Terminal Slots)	Rate Of Station Response	Prob. Station Busy	% Of IP Blocked	Total % Of IP Loss	Terminals Remaining
CCSDR	13	12.14	16.61	2.06	.50	10.63	11.41	10
	25	12.20	34.97	1.61	.48	29.50	32.95	23
CCTDR	13	12.39	4.91	1.99	.26	1.59	1.59	9
	25	23.33	11.51	1.97	.31	3.31	3.31	34

estimating the maximum throughput. Figure 14.8 shows the throughput versus offered rate for CCSDR and CCTDR systems. The curves are linear for low offered rates and saturate when the offered rate increases.

For the CCSDR system one can see that the throughput is practically the same when the offered rate is increased from 13% to 25%. This and the other measures (see Table 14.2), (for example, the rate of station response) show that the system is overloaded at a 25% offered rate. On the other hand, the system seems to operate at steady state at an offered rate of 13% (rate of station response 2.06). A rough estimate of maximum throughput for this system would be between 12% and 15%. Similar observations of the performance measures lead to an "estimate" of between 27% and 30% for the maximum throughput of the CCTDR system in the specified repeater configuration.

The average delay of the first Information Packet from terminal to station, and the Total Loss, as a function of offered rate are shown in Figure 14.9 and Figure 14.10 respectively.

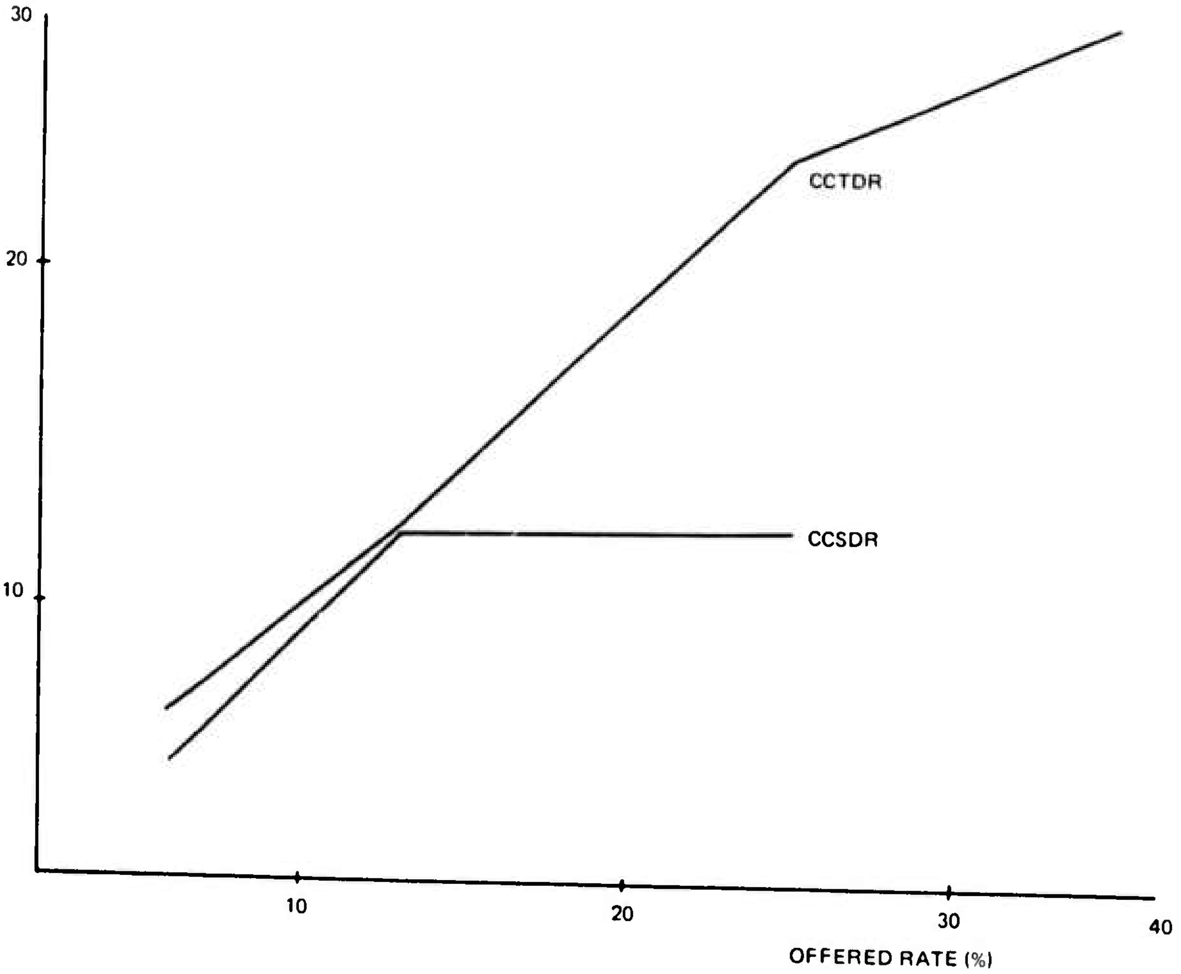


Figure 14.8: System Throughput vs. Offered Rate

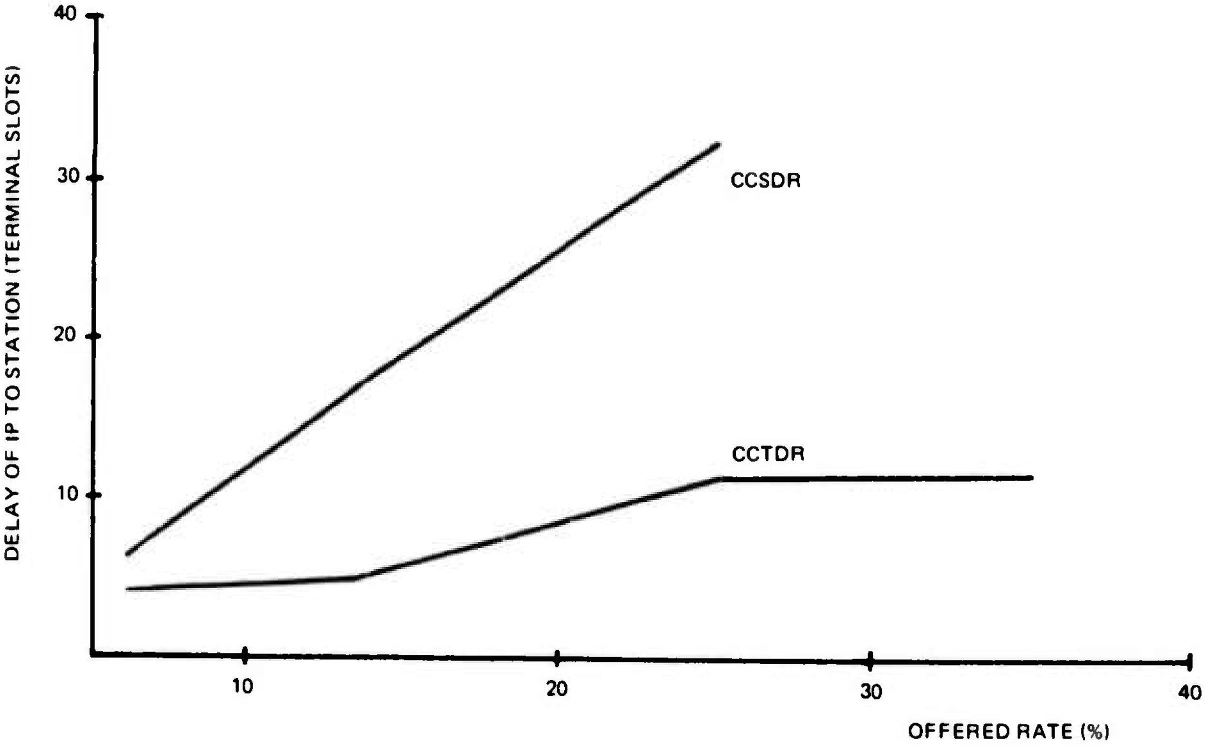


Figure 14.9: Terminal-Station Delay vs. Offered Rate

REMARKS: There are many parameters in the simulation program which we have not experimented with (or tried to optimize) and which affect the quantities discussed above. One parameter which is significant in determining the maximum throughput is the average number of response packets from station to terminal. The effect of this parameter has been analyzed in [28], for a slotted ALOHA random access mode. It has been shown that the maximum throughput is increased in the Common Channel system when the rate of response increases, and the maximum throughput tends to 100% of the data rate when the rate of response tends to infinity. We expect that this parameter has a similar effect for the mode of access simulated. In the results reported here the rate of response is 2.0 which is small compared to usual estimates for terminals interacting with computers. Furthermore, the relatively short terminal interaction increases the traffic overhead of the search procedure per information packet.

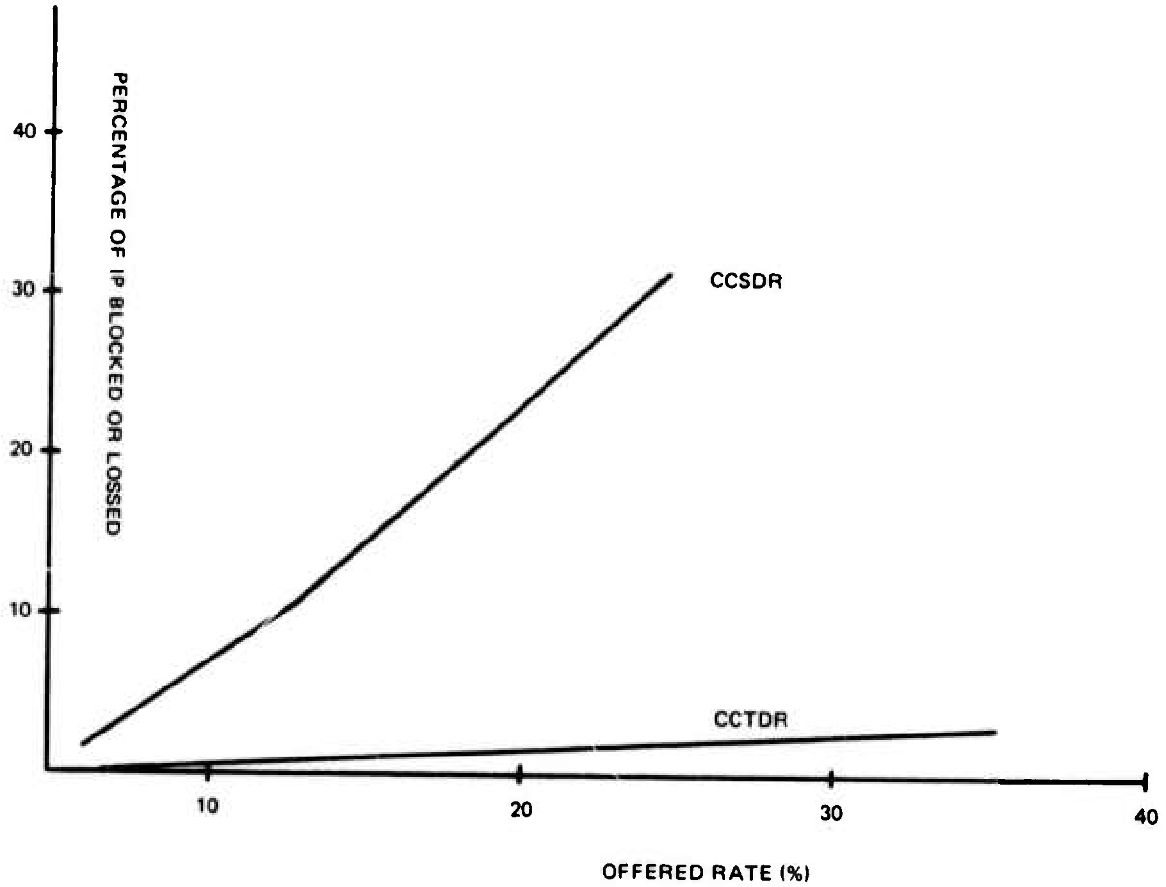


Figure 14.10: IP Blocking vs. Offered Rate

Chapter 15
PACKET RADIO SYSTEM AREAS FOR FURTHER STUDY

As is evident, packet radio offers a new and challenging area for network analysis and design. Previous studies have merely touched crucial areas. Further study to develop methodology, to provide support for hardware and software design, and to effectively control and manage network resources is required. A variety of studies is currently underway. These studies will:

a. Estimate system capacity as a function of terminal-repeater and repeater-repeater signaling rates for multistation networks.

b. Compare the performance of systems with varying degrees of receiver capture, multiple channels, directional antennas and multiple detectors.

c. Determine efficient operating parameters including time out intervals, handover numbers, and number of retransmissions.

d. Determine relationship between number of repeaters, throughput, delay and blocking.

e. Compare the efficiency of direct terminal to terminal routing versus hierarchical routing in multistation networks.

f. Estimate throughput, delay, and blocking for multistation networks.

g. Develop and test multistation algorithms for routing and labeling and relabeling.

h. Develop high level global flow control algorithms to allow effective utilization of system resources.

i. Determine network control strategies to identify and monitor network congestion and element failure conditions.

j. Formulate dynamic reliability and survivability criteria and develop algorithms for network reliability analysis and design.

k. Develop algorithms for configuring packet radio networks to meet specified reliability and survivability criteria.

REFERENCES

1. Abramson, N., "The ALOHA System—Another Alternative for Computer Communications," *AFIPS Conference Proceedings,* Vol. 37, November, 1970, pp. 281-285.

2. Abramson, N., "Packet-Switching with Satellites," *National Computer Conference,* June, 1973, pp. 695-702.

3. Boehm, S. P., P. Baran, "Digital Simulation of Hot-Potato Routing in a Broadband Distribution Communications Network," Rand Corporation, *Memorandum RM-3103-PR,* August, 1964.

4. Chou, W., and H. Frank, "Routing Strategies For Computer Network Design," *Proceedings of Brooklyn Polytechnic Institute on Computer Networks and Teletraffic,* April, 1972.

5. Chou, W., M. Gerla, and H. Frank, "Communication Network Cost Reduction Using Domestic Satellites," *Symposium on Computer Networks: Trends and Applications,* National Bureau of Standards, Gaithersburg, Maryland, May, 1974.

6. Cox. J. E., "Western Union Digital Services," *Proceedings of the IEEE,* Vol. 60, No. 11, November, 1972, pp. 1350-1356.

7. Durfee, E. W., and R. T. Callais, "The Subscriber Response System," *National Cable Television Association Convention Record,* Washington, D.C., 1971, pp. 28-48.

8. Eldridge, F. R., "System for Automatic Reading of Utility Meters," *MITRE Report,* M72-7, the MITRE Corporation September, 1971.

9. Fralick, S.C., "R. F. Channel Capacity Considerations," available from *ARPA Network Information Center,* Stanford Research Institute, Menlo Park, Calif., 1974.

10. Frank, H., and W. Chou, "Throughput In Computer-Communication Networks," In *Infotech Report on The State Of The Art of Computer Networks,* 1972.

11. Frank, H., I. T. Frisch, and W. Chou, "Topological Considerations In The Design of The ARPA Computer Network," *In Spring Joint Computer Conference Proceedings*, Washington, D.C., Spartan, 1970, pp. 581-587.

12. Frank, H., and W. Chou, "Topological Optimization of Computer Networks," *Proceedings of the IEEE*, Vol. 60, No. 11, pp. 1385-1396.

13. Frank, H., and W. Chou, "Routing in Computer Networks," *Networks*, Vol. 1, No. 2, pp. 99-112.

14. Frank, H., and I. T. Frisch, "Analysis and Design of Survivable Networks," *IEEE Transactions on Communication Technology*, Vol. COM-8, October, 1970, pp. 501-519.

15. Frank, H., and I. T. Frisch, *Communication, Transmission, and Transportation Networks*, Addison-Wesley, Reading, Mass., 1971.

16. Frank, H. and I. T. Frisch, "The Design of Large Scale Networks," *Proceedings of the IEEE*, Vol. 60, No. 1, January, 1972, pp. 6-11.

17. Frank, H., and I. T. Frisch, "Planning Computer-Communication Networks," *Computer-Communication Networks*, Eds. N. Abramson and F. F. Kuo, Prentice-Hall, Englewood, N. J., 1973, pp. 1-28.

18. Frank, H., R. Kahn, and L. Kleinrock, "Computer Communication Network Design—Experience with Theory and Practice," in *Spring Joint Computer Conference, AFIPS Conference Proceedings*, Washington, D.C.: Spartan, 1972.

19. Fratta, L., M. Gerla, and L. Kleinrock, "The Flow Deviation Method: An Approach to Store-and-Forward Communication Network Design," *Networks*, Vol. 3, No. 2, pp. 97-134.

20. Frisch, I. T., "Technical Problems in Nationwide Networking and Interconnection," *IEEE Transactions on Communications*, January, 1975.

21. Frisch, I. T., B. Rothfarb, and A. Kershenbaum, "A Computer Design of CATV Distribution Systems," *Cablecasting*, Vol. 7, July-August, 1971, pp. 20-26.

22. Fuchs, E., and P. E. Jackson, "Estimates of Distributions of Random Variables for Certain Computer Communications Traffic Models," *Communications of the ACM*, Vol. 13, N12 December, 1970, pp. 752-757.

23. Fulkerson, D., G. Nemhauser, and L. Trotter, "Two Computationally Difficult Set Covering Problems that Arise in Computing in 1-Width of Incidence Matrices of Steiner Triple Systems," *Technical Report 903*, Cornell University, Department of Operations Research, 1973.

24. Gabriel, R. P., "Dial A Program—An HF Remote Selection Cable Television System," *Proceedings of the IEEE,* Vol. 58, No. 7, July, 1970, pp. 1016-1023.

25. Gaines, E. C., "Specialized Common Carriers—Competition and Alternative" *Telecommunications,* September, 1973, pp. 17-26.

26. Garfinkel, R. and G. Nemhauser, *Integer Programming,* J. Wiley, New York, 1972.

27. Gerla, M., W. Chou, and H. Frank, "Cost-Throughput Trends in Computer Networks Using Satellite Communications," *International Conference on Communications,* ICC-74, June 17-19, Minneapolis, Minnesota, pp. 31C-1-21C-5.

28. Gitman, I., R. M. Van Slyke and H. Frank, "On Splitting Random Access Broadcast Communication Channels," *Proceedings of the Seventh Hawaii International Conference on System Sciences,* Subconference on Computer Nets, January, 1974.

29. Gitman, I., "On the Capacity of Slotted ALOHA Network and Some Design Problems," *IEEE Transactions on Communications,* March, 1975.

30. Gross, W. B., "Distribution of Electronic Mail Over the Broadband Party-Line Communications Network," *Proceedings of the IEEE,* Vol. 58, No. 7, July, 1970, pp. 1002-1012.

31. Hayes, J. F., and D. N. Sherman, "Traffic Considerations of a Ring Switched Data Transmission System," *Bell Systems Technical Journal,* Vol. 50, No. 9, November, 1971, pp. 2947-2978.

32. Hayes, J. F., and D. N. Sherman, "A Study of Multiplexing Techniques and Delay Performance," *Bell System Technical* Journal, Vol. 51, No. 9, November, 1972, pp. 1983-2010.

33. Jackson, P. E. and C. D. Stubbs, "A Study of Milti-access Computer Communications," *Proceedings AFIPS 1969 Spring Joint Computer Conference,* Vol. 34, AFIPS Press, Montvale, New Jersey, pp. 491-504.

34. James, R. J., and P. E. Muench, "AT&T Facilities and Services," *Proceedings of the IEEE,* Vol. 60, No. 11, November, 1972, pp. 1342-1349.

35. Jerrold Electronics Corporation, 1971 National Cable Television Convention Publicity Release on Two-Way CATV Systems.

36. Kirk, D., and M. J. Paolini, "A Digital Video System for the CATV Industry," *Proceedings of the IEEE,* Vol. 58, No. 7, July, 1970, pp. 1026-1035.

37. Kershenbaum, A., and R. M. Van Slyke, "Recursive Analysis of *Networks*," Vol. 3, No. 2, 1973, pp. 81-94.

38. Kleinrock, L., and F. Tobagi, "Carrier Sense Multiple Access for Packet-Switched Radio Channels," *IEEE International Conference on Communications*, Minneapolis, Minnesota, June, 1974.

39. Lancaster, P. W., and J. Garodnik, "CATV Environment for Data Communication," *National Telecommunications Conference*, Atlanta, November, 1973, pp. 38C-1-38C-4.

40. Mertz, P., "Influence of Echoes on TV Transmission," *Journal of the SMPTE*, May 1953.

41. NAC (Network Analysis Corporation), "The Practical Impact on Recent Computer Advances on the Analysis and Design of Large Scale Networks," *Third Semiannual Technical Report*, June, 1974.

42. Okamura, Y. et. al., "Field Strength and Its Variability on UHF and UHF Lan-Mobile Radio Service," *Review of the Electrical Communication Laboratory*, Vol. 16, No. 9-10, September-October, 1968.

43. Olszewski, J. A., and H. Lubars, "Structural Return Loss Phenomenon in Coaxial Cables," *Proceedings, of the IEEE*, Vol. 58, No. 7, July, 1970, pp. 1036-1050.

44. Ornstein, S. M., F. E. Heart, W. R. Crowther, H. K. Rising, S. B. Russell, and A. Michel, "The Terminal IMP for the ARPA Computer Network," *Proceedings AFIPS 1972 Spring Joint Computer Conference*, Vol. 40, AFIPS Press, Montvale, New Jersey, pp. 243-254.

45. Reinfelder, W. A., *CATV System Engineering* TAB Books, Blue Ridge Summit, Pennsylvania, 1970.

46. Roberts, L. G., "ALOHA Packet System With and Without Slots and Capture," *ARPANET Satellite System Notes 8*, (NIC Document #11291), available from ARPA Network Information Center, Stanford Research Institute, Menlo Park, California, June, 1972.

47. Rothfarb, B., and M. Goldstein, "The One Terminal Telpak Problem," *Operations Research*, Vol. 19, No. 1, February, 1971, pp. 156-169.

48. Rogeness, G. "Contributing Sources and Magnitudes of Envelope Delay in Cable Transmission System Components," *National Cable Television Association Convention Record*, Chicago, May, 14-17, 1972, pp. 479-506.

49. Schwartz, M., W. R. Bennett, and S. Stein, *Communication Systems and Techniques*, McGraw-Hill, 1966.

50. Schwarz, M., R. R. Boorstyn, and R. L. Pickholtz, "Terminal-Oriented Computer Networks," *Proceedings of the IEEE*, Vol. 60, No. 11, November, 1972, pp. 1408-1422.

51. Switzer, I., "The Cable System as a Computer Network," *Proceedings of Symposium: Computer-Communications Networks and Teletraffic*, Polytechnic Institute of Brooklyn, New York, April, 1972, pp. 339-346.

52. Turin, G. L., "A Statistical Model of Urban Multipath Propagation," *IEEE Transactions on Vehicular Technology*, Vol. VT-21, February, 1972, pp. 1-9.

53 Transmission Systems for Communications, Bell Telephone Laboratories, 1971.

54. Van Slyke, R., and H. Frank, "Network Reliability Analysis I, *Networks*, Vol. I, No. 3, 1972, pp. 279-290.

55. Van Slyke, R., and H. Frank, "Reliability of Computer-Communication Networks," *In the Proceedings of the 5th Conference Application of Simulation*, (New York, N.Y.), December, 1971.

56. Volk, J., "The Reston Virginia Test of the MITRE Corporation's Interactive Television System," MITRE Corporation, May, 1971.

57. Ward, J. E., "Present and Probable CATV/Broadband Communication Technology," Sloan Commission Report on Cable Communications, On the Cable: *The Television of Abundance*, McGraw-Hill, 1971.

58. Willard, D. G., "MITRIX: A Sophisticated Digital Cable Communications System," *National Telecommunications Conference*, Atlanta, November, 1973, pp. 38E-1-38E-5.

59. Worley, A. R., "The Datran System," *Proceedings of the IEEE*, Vol. 60, No. 11, November, 1972, pp. 1357-1368.

60. Yaged, B., Jr., "Minimum Cost Routing for Dynamic Network Models," *Networks*, Vol. 3, No. 3, pp. 193-224.

part of the signal as well, and thus is area dependent. For W, this means one can shape the critical region somewhat asymmetrically now in the complex frequency plane in order to accommodate the jittery signal source and its Rayleigh noise vector over the newly expanded region defining the location of the signal source. If the signal source is distributed uniformly or nearly so, or when the SNR is high, then envelope detection is to be preferred.

Having broached the concept of an asymmetrical critical region, one sees that better pde's for partially coherent systems could be devised if one fitted anomalous critical regions to the two-dimensional pdf of the varying signal source. The present use of W represents but a tentative step in this direction largely initiated by the comparative ease with which the parameters of the system can be calculated and the simple techniques required for measuring the important aspects of the signal.

The numerical examples cited here indicate that the W statistic could be an option for use in situations involving dispersive media (where signal shape tends to be lost causing the operation of matched filters to be degraded), when a low to moderate SNR exists (when envelope detection is at a disadvantage) and when mild phase fluctuations exist (causing synchronous detection to be impaired).

There is a second property of W which might be noted, though, perhaps, it is not too important. That is, incoherent and synchronous detection can be treated as two special cases of W: when k is equal to 0 and the phase is uniformly random, and when the phase is known and k is equal to $y/2^{1/2} \cdot Z$, respectively. This second property is of value largely as a conceptual unifier, but the former quality, the ability for handling variable phase through the use of an irregular two-dimensional critical region, and the possibilities it suggests, should prove of interest to workers in communications, radar and control systems.

REFERENCES

[1] M. Schwartz, W. R. Bennett, and S. Stein, *Communication Systems and Techniques.* New York: McGraw-Hill, 1966.

[2] S. O. Rice, "Mathematical analysis of random noise," *Bell Syst. Tech. J.*, vol. 24, pp. 46–157, 1945.

[3] E. D. Sunde, *Communication Systems Engineering Theory.* New York: Wiley, 1969.

[4] F. S. Weinstein, "Simplified relationships for the probability distribution of the phase of a sine wave in narrow-band normal noise," *IEEE Trans. Inform. Theory* (Corresp.), vol. IT-20, pp. 658–661, Sept. 1974.

Routing in Packet-Switching Broadcast Radio Networks

I. GITMAN, MEMBER, IEEE,
R. M. VAN SLYKE, MEMBER, IEEE, AND
H. FRANK, MEMBER, IEEE

Abstract—**Packet-switching broadcast radio networks are receiving considerable attention as a feasible solution for applications involving fast network deployment requirements, inaccessible physical environments, and mobile communication devices. Such networks also offer economic alternatives to traditional multiplexing schemes for local distribution.**

Most of the published papers relating to packet-switching broadcast radio networks address the case in which all communication devices are within an effective transmission range of the destination receiver, thus forming a single-hop network in which no packet routing is involved. In this paper, we address multihop networks. The problems encountered in packet transportation are identified and strategies to resolve these are proposed.

Paper approved by the Editor for Computer Communication of the IEEE Communications Society for publication without oral presentation. Manuscript received August 5, 1975; revised February 17, 1976. This work was supported by the Advanced Research Projects Agency of the Department of Defense under Contract DAHC-15-73-C0135.

The authors are with the Network Analysis Corporation, Glen Cove, NY 11542.

I. INTRODUCTION

Consider a set of resources which must exchange messages via a packet communication network. Network tasks related to packet transportation can be classified as follows: 1) origination and destination functions; 2) relay functions; 3) management and control functions; and 4) gateway functions. The gateway functions are needed only when one considers message transportation between nonhomogeneous networks (i.e., when the origination and destination resources do not "reside" in the same network). The differences between the packet-switching broadcast network we consider and point-to-point networks (such as ARPANET) include the following.

1) A communication channel (i, j) in a point-to-point packet-switching network is associated with its end nodes i and j. Thus, node i can transmit to node j on channel (i, j) without addressing the packet to node j. On the other hand, a channel in a radio network cannot be associated with two nodes. Packet transmissions (we assume omnidirectional antennas) may be received by all nodes within range of the node that emitted the packet. If node i wishes to send a packet to node j (only), it must add to the packet header an instruction that all other nodes should discard the packet. This also implies that if a node receives a packet with an error, it may not know that the packet was addressed to it and consequently, cannot request a retransmission.

2) In a broadcast radio network, channels are shared by sets of nodes and cannot be dedicated to specific pairs. Time can be divided into nonoverlapping intervals and assigned to each node. However, numerous studies (e.g., [1], [2], [9], [10], [15], [6], [7]) have shown that this "fixed capacity assignment" is wasteful for many applications. We shall therefore assume that nodes use a "random transmission scheme" (see above references) which results in a dynamic sharing of the channel capacity without centralized control. (The specific scheme used is of no significance to the issues addressed here.) This implies that a radio node can simultaneously receive several packets, all in error. Hence, the probability of receiving a packet with error is much greater than on a point-to-point channel, and varies as a function of traffic level, and the spatial distribution of traffic sources and nodes.

Packet radio networks are particularly suitable for applications in which: 1) resources (e.g., terminals, computers) are mobile, so that a broadcast mode is necessary; 2) resources are located in remote or hostile locations where hardwire connections are uneconomical or not feasible; and/or 3) the traffic characteristics of resources are of a bursty nature; that is, there is a high ratio of peak bandwidth to average bandwidth requirements. Specific packet radio networks in the stage of design or in current use were discussed in [14].

A network is a "single-hop" network if no relay functions are needed at its nodes. Most analyses of radio networks address single-hop networks (e.g., [2], [9]). Packet routing in radio networks was first addressed in [12], [13], [3], [8]. The analysis of multihop packet radio networks is extremely difficult. Consequently, one must use simulation [13] or study simple models to identify network properties. Without efficient routing and flow control in large scale radio networks: 1) a packet may endlessly circulate among nodes;

2) many copies of a packet will be generated; and 3) if the destination node is on another network, many copies of a packet may arrive at different gateway nodes and be introduced into other networks.

Idealized results of packet proliferation in radio networks derived from analyses of simplified combinatorial models [12] have indicated that without proper controls, the number of copies generated by a single packet may grow exponentially. Thus, an improperly designed broadcast radio network can easily be saturated and cease to fulfill its function. Consequently, the routing and flow control stategies used for packet transportation are of utmost importance for efficient network operation.

The network model we examine contains a large number of nodes using random transmission schemes and omnidirectional antennas. The network contains two types of nodes. A node with origination, destination, and relay functions is called a repeater. A node with additional capabilities such as gateway, global control, global initialization, accounting, and directory functions, is called a station. To cover a large geographical area with a communication network to serve mobile terminals, not all nodes must perform all functions. Repeaters with limited capabilities can provide area coverage. Stations are assigned to satisfy capacity and reliability requirements. In this paper, we do not discuss terminals, users, or other network resources. These elements are not an integral part of the *communication* network and do not directly affect the problems discussed here.

The problems addressed are: flow control functions related to packet transportation, the rules for packet transportation (routing strategies), and the network architecture implied by the routing. The objectives of packet transportation are: 1) reliability: to assure that a packet launched into the network will reach its destination with high probability; and 2) efficiency: to deliver a large number of messages with a relatively small time delay. In general, given two routing algorithms A and B, we say that A is more efficient than B if it uses less network resources (channel and node capacities per packet) than B does.

We examine three approaches to routing in radio networks: a broadcast routing algorithm which satisfies objective 1) but fails to satisfy objective 2); and two algorithms which satisfy both objectives for different traffic patterns. The broadcast algorithm may be needed in an operational network as a backup algorithm or for initialization of nodes to use more efficient algorithms.

II. BROADCAST ROUTING ALGORITHM

In point-to-point networks, the routing algorithm must determine outgoing lines for packets. In radio networks, the major decision is not to determine the next node, but to either accept a packet for switching or reject it. The broadcast routing algorithm is essentially a flow control procedure which prevents looping and cycling of packets. It contains the following mechanisms.

1) A hop-by-hop acknowledgement to guarantee that the packet is accepted by the next repeater.

2) A (maximum) handover number (carried in the packet), which is decremented by a repeater which accepts the packet for switching. This guarantees that the packet will traverse no more than the maximum number of hops assigned.

3) A variable transmission power mechanism that can increase power (and hence the number of potential receivers) as a function of the number of transmissions without acknowledgement.

4) Random transmission control parameters such as time interval for rescheduling unacknowledged packets and maximum number of transmissions per packet.

5) A time parameter FORGET, which is the maximum interval of time during which a packet previously switched by a repeater will not be accepted for switching.

The reader may conceive of several algorithms which utilize variations of items 1) through 5). Items 1) through 4) are controls recommended for all radio routing algorithms. These do not prevent looping and cycling of packets. The mechanism which defined the broadcast routing algorithm is item 5).

To implement broadcast routing, every packet has a unique identifier. Repeaters have storage for L such identifiers. When a packet is acknowledged or discarded, its identifier and the time are recorded by the repeater. When a new packet is received, its identifier is compared with stored identifiers. If a match occurs, the packet is discarded. Thus, if a repeater has available storage, it accepts a packet if it did not switch the same for at least FORGET seconds or it switched at least L other packets since the previous switching of the received packet.

A. Properties

1) The algorithm is nondirectional with no addressing along the path. A destination node recognizes packets sent to it by comparing the destination ID against its own ID.

2) If L and FORGET are large, a packet will not be switched by a node more than once.

3) The algorithm is simple and reliable. Repeaters need not know network connectivity or destination node locations. If a route exists and the maximum handover number is large compared to the shortest path to the destination, packets will reach their destinations with high probability.

4) The algorithm inefficiently utilizes network resources, a large number of duplicate packets may be generated. Moreover, when a "radio" packet is destined for another network, copies of the packet may be introduced into the other network by several gateway nodes unless prevented by gateway communication protocols.

B. Discussion

Radio network features not possible in point-to-point networks are transmission power control and a special form of hop-by-hop acknowledgment (HBH ack). Transmission power control enables bypass of failed nodes to increase network reliability. Acknowledgments need not be explicitly transmitted since subsequent retransmissions by the receiving node are also received by the previous transmitting node. The latter node recognizes the ack by comparing the identifier of the packet received and its handover number against those of packets waiting for retransmission.[1] This is more efficient than a specific acknowledgement since a single packet transmission can acknowledge several nodes.

The broadcast routing algorithm floods either the network or a subset of nodes. Flooding can be utilized by stations to map network connectivity so that routing information can be assigned to repeaters to obtain more efficient routing. It also enables a station to change parameters in all repeaters. Finally, broadcast routing can be used in networks whose nodes are mobile such that updating routing information based on connectivity becomes infeasible.

III. HIERARCHICAL ROUTING ALGORITHMS

The primary shortcoming of the broadcast routing algorithm is its nondirectionality. This limitation is addressed by the routing procedures discussed below.

[1] This form of HBH ack assumes that if node i can receive from node j, node j can also receive from node i. We assume that every node has at least one such link, otherwise it is considered disconnected.

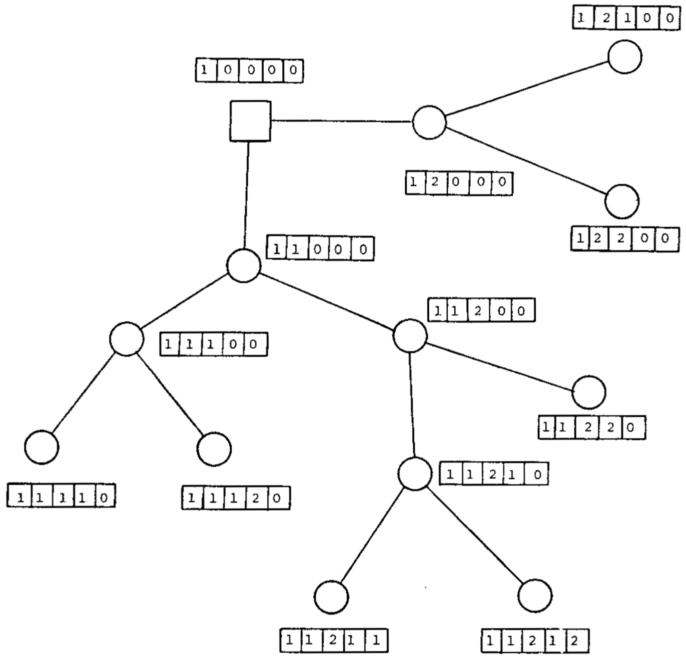


Fig. 1. Hierarchical labels of repeaters and stations, and the tree structure formed.

A major use of the radio network will be local collection and distribution of traffic. This traffic primarily flows from repeaters to stations and from stations to repeaters. If traffic is to be routed between repeaters, we can insist that it pass through a station creating a hierarchical routing requirement. Among other virtues, this requirement allows centralized control.

The techniques suggested assign, during initialization, routing information to repeaters. The information assigned by a station is called a "label." The set of repeater labels forms a hierarchical tree structure of repeaters rooted at the station. Labels may be changed during network operation when changes in network topology occur.

The routing strategy identifies shortest path (minimum hop) between repeaters and stations, and prevents, wherever possible, generation of duplicate copies of packets. The routing is sufficiently flexible to allow departures from the first choice path and use of next shortest paths.

A. The Hierarchical Label for Single-Station Networks

Repeater lables form a hierarchical structure as shown in Fig. 1. Each label includes the following information: 1) routing address of the repeater, 2) minimum number of hops to the station, and 3) names in a special compact representation of all repeaters on a shortest path to the station. In particular, the address of the next repeater along the transmission path is readily available.

A label is composed of H fields. The label of repeater R at a distance of $j - 1$ hops to the station contains nonzero integers for the first j fields and zeros for the remaining $H - j$ fields. We say that R is at level j of the hierarchy. The repeater to which R addresses its packets when routing towards the station is called the "home" of R. The labels of a set of repeaters at level j, which have the same home repeater, differ only in the entry in field j. Thus, the label of the station has a nonzero entry in the first field and a zero in all other fields; the labels of repeaters at a distance of one hop to that station have the station's entry in the first field, nonzero unique entries in the second field, and zero in all other fields, etc. Fig. 1 shows an example of a labeled set of repeaters.

B. The Routing Algorithm

The complete path between the station and a repeater is defined by the label of the repeater. A handover number is used for flow control and for the HBH ack test as in the broadcast routing algorithm. An ALL indicator is used for alternate routing. When the ALL indicator is not active, the packet is addressed to a single repeater with an address defined by the hierarchy level indicator and the label. The hierarchy level indicator is a pointer to the label and defines the number of nonzero fields of the repeater; if these nonzero fields match those in the packet label, the packet is addressed to this repeater.

When a repeater exceeds its allowed number of retransmissions without receiving an HBH ack, it begins alternate routing by activating the ALL indicator in the packet. Thereafter, all repeaters which match the hierarchy level indicator may switch the packet. These new repeaters use the same routing address and thus attempt to regain the primary route. Thus, a failed repeater or temporary busy repeater may be bypassed. Fig. 2 shows schematically the packet flow when using alternate routing.

C. Routing in Multistation Radio Networks

In a multistation network, repeaters are assigned labels by several stations during initialization. Repeaters determine primary label, secondary label, etc., according to distance, in number of hops, to the corresponding stations. Order may be changed upon changes in network topology.

A repeater can route packets to any station via the single-station algorithm. A repeater matches the packet label with one of its labels and then decrements or increments the hierarchy level indicator depending on whether the packet is directed TO or FROM a station. A two-station example for the radio network shown in Fig. 3 is given in Fig. 4. Fig. 4 shows the partition of the set of repeaters between the two stations generated by the choice of the primary and secondary labels. All repeaters above the separation line have a smaller number of hops to $S - 1$, while repeaters below the line have chosen $S - 2$ as their primary station. $R4$, $R11$, and $R18$ have the same number of hops to the two stations and their choice is arbitrary.

D. Properties

1) The algorithm allows shortest path routing between repeaters and stations when the labels are properly assigned. This results in less utilization of nodal processing capacity and better use of channel capacity. Thus, average delay experienced when using the hierarchical algorithm will be lower than for the broadcast algorithm because only repeaters on the path transmit the packet. This yields smaller interference probability, fewer transmissions per hop before success, and consequently a smaller delay per hop.

2) Packets will reach their destinations with high probability. Since routing is directional, packets will arrive at only one station.

3) The algorithm requires maintenance of an updated set of labels. This becomes a limitation only in a network with highly mobile nodes.

E. Discussion

Two of the possible generalizations of the hierarchical routing algorithm are discussed below.

Generalization 1: Transmission power is increased when blocking is encountered, instead of the ALL indicator. All repeaters on the preferred path that are closer to the destination can accept the packet. The advantage of this technique is that one can shorten the number of hops to the destination and bypass failed repeaters. The limitations are: 1) duplicate

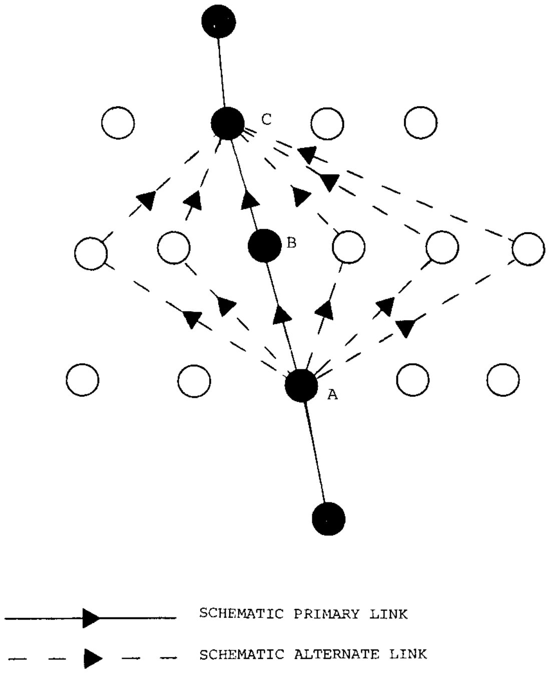


Fig. 2. Alternate routing example.

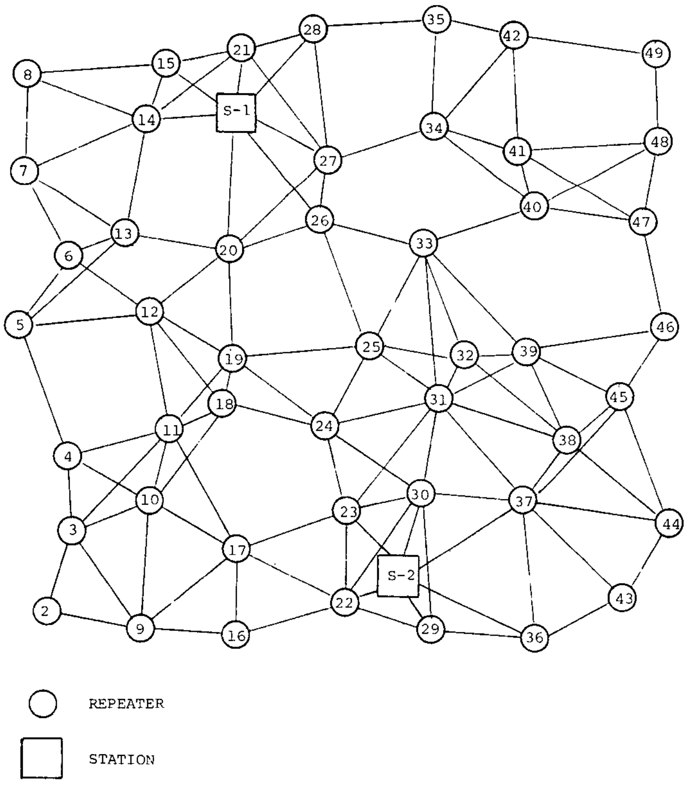


Fig. 3. Connectivity of a radio network with two stations.

copies of the packet may be generated; 2) increase in power results in interference with a large number of repeaters; and 3) difficulties may be encountered in the HBH ack scheme when the receiving repeater transmits the packet with a lower power than that used by its predecessor.

Generalization 2: This generalization involves the search for an alternate repeater when encountering blocking. A "search" protocol uses a search packet specifying the class of repeaters sought. A response by such a repeater contains the hierarchical label of the responding repeater. The searching repeater substitutes the label of the responding repeater into the original packet and transmits the packet. The class of repeaters sought can be progressively increased as necessary.

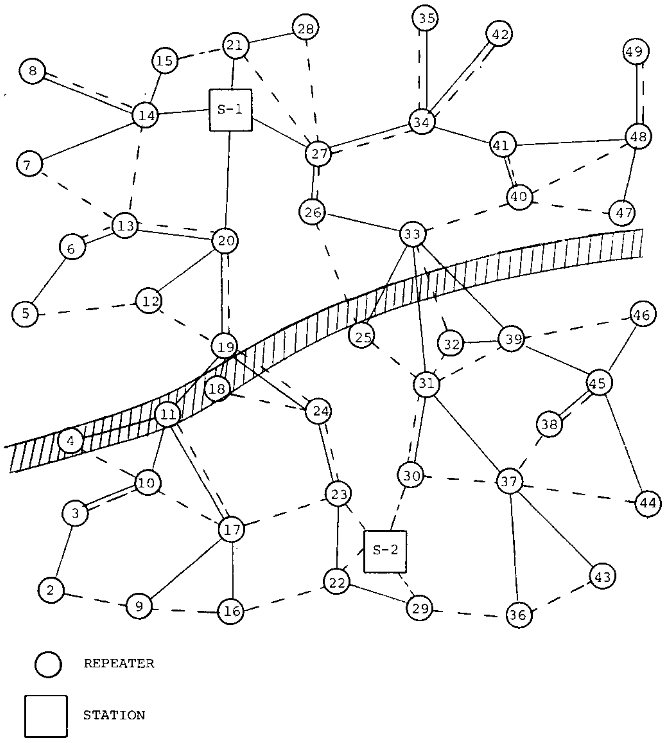


Fig. 4. Partition of repeaters to primary and secondary station.

With this technique no duplicate copies of the packet are generated along the path since the packet is always transmitted to a unique repeater. However, two hierarchical labels in the packet header are required when routing from the station. Moreover, the search protocol must be implemented within the repeater, which requires additional memory.

IV. DIRECTED BROADCAST ROUTING ALGORITHM

The hierarchical routing algorithm is most useful when the radio network is used for local collection and distribution of traffic. Furthermore, if stations contain directories and accounting capabilities, every packet must pass through a station and thus no penalty is associated with routing all traffic through a station. We now suppose that: 1) traffic flows are equally likely between any pair of nodes, 2) the traffic origination device knows (or can obtain) the ID of the destination repeater, and 3) that accounting per packet or per message is not needed. With these assumptions, a distributed network architecture is appropriate. Hierarchical routing can still be used but distributed routing whereby a packet can be routed directly to the destination repeater may be more efficient.

We still assume that repeaters have limited capabilities, and that stations perform all the tasks not directly related to packet switching. In a distributed radio network, stations act mainly as observers and controllers performing: 1) network initialization, mapping network topology, determining labels, and initializing repeaters; 2) monitoring network connectivity and changing repeater labels; 3) global flow control functions and controlling operating parameters of repeaters; 4) gateway functions (traffic to destinations not in the radio network must pass through a station); and 5) supervisor (extra capabilities, such as resource directories, required by terminals or repeaters will be supplied by a station).

A. The Repeater Label

The label contains similar information to ARPANET IMP's [4], [5]. There, node i has a table A of $(N-1) \times L_i$ entries, where N is the number of nodes in the network and L_i is the number of outgoing point-to-point channels from node i. An entry a_{kj} indicates the distance (or delay) from node i to node k when using outgoing link j.

In the radio system, a repeater is assigned a label in the form of a distance vector. Repeater R_i will be given the vector $\boldsymbol{d}_i = (d_{ij})$, where d_{ij} is the minimum number of hops from R_i to R_j. If there are N nodes in the network (repeaters and stations), the maximum distance is $N - 1$ hops, and we also must represent a distance of ∞.[2] Thus, one needs a total of $N \log_2 N$ bits for storing the vector $\boldsymbol{d}_i$. Note that R_i is not provided with the ID's of its neighbors. Its own ID (R_i) is not explicitly known to it, but implicitly given by the fact that $d_{ii} = 0$. (Its neighbors are implicitly known by $d_{ij} = 1$.)

B. The Routing Algorithm

The destination ID and the distance to it from the transmitting device are used for the HBH ack and to determine the devices that should accept a packet for switching. In this case, the packet acceptance rule is that the receiving device is nearer to the destination than the transmitting device. The destination device identifies packets for it by noticing that its distance to the destination is zero.

When encountering blocking, a repeater may increase its transmission power. This may result in some difficulties in obtaining an HBH ack, as discussed in Section III-E. As a last option, the repeater may send a control packet to a station to resolve the difficulty.

C. Properties

1) The algorithm enables direct (i.e., not via a station) routing to any repeater or station in the radio network.

2) There is a high probability of reaching the destination node. Ignoring traffic overload, if there is a path to the destination, the packet will "find" it. The packet is routed to a single destination.

3) The algorithm uses shortest path routing because the acceptance scheme eliminates repeaters on the basis of distance to the destination, rather than the number of hops traversed.

4) The algorithm allows generation of copies of the packet along the path. Thus, its efficiency will be less than hierarchical routing for repeater-to-station paths.

5) The main limitation is the need to maintain an up-dated set of labels. The algorithm as presented is suitable for stationary network nodes. If the application implies a distributed architecture and the repeaters and stations are mobile, the task of updating distance vectors should be distributed. This will require more repeater capabilities.

V. CONCLUDING REMARKS

Three routing approaches for broadcast radio networks were proposed in this paper. The approaches take advantage of the properties of radio networks to minimize the utilization of network resources for packet transportation.

The broadcast routing algorithm is the least efficient; however, devices need not be initialized. The comparison between the hierarchical routing algorithm and the directed broadcast algorithm is not obvious. The directed broadcast algorithm seems to be more reliable because more than one receiver may accept a packet for switching. Reliability is further enhanced if an adaptive power mechanism is available in repeaters. Also, the HBH ack scheme should cancel out many of the possible paths, especially as a packet nears its destination. However, on a path between a repeater and a station, the directed broadcast algorithm will utilize more netwrok resources than the hierarchical routing algorithm.

Other potentially major points for comparison are flexibility in network expansion and the overhead of updating repeater labels. Here, the hierarchical routing algorithm is preferable to the directed broadcast algorithm. When a new repeater is added, all repeaters' labels must be changed for the directed broadcast technique. With the hierarchical algorithm, it is only necessary to initialize the new repeater. Furthermore, the fact that all packets are routed via a station, in the hierarchical algorithm, enables easy monitoring of network connectivity and faster updating of labels.

An analytic comparison of the algorithms is difficult. The broadcast and hierarchical algorithms were simulated and extensively tested, some results appear in [3]; the directed broadcast algorithm has not yet been experimentally evaluated.

[2] The distance of ∞ can be used by a station to prevent communication between a repeater and a set of repeaters, for partitioning the repeater network among the stations, or possibly for turning a repeater off, when all entries apart from d_{ii} are ∞.

REFERENCES

[1] N. Abramson, "The ALOHA system—Another alternative for computer communication," in *1970 Fall Joint Comput. Conf., AFIPS Conf. Proc.*, vol. 37. Montvale, NJ: AFIPS Press, 1970, pp. 281–285.

[2] N. Abramson, "Packet switching with satellites," in *Proc. Nat. Comput. Conf.*, June 1973, pp. 695–702.

[3] H. Frank, I. Gitman, and R. Van Slyke, "Packet ratio system—Network considerations," in *Proc. Nat. Comput. Conf.*, May 1975, pp. 217–231.

[4] G. L. Fultz and L. Kleinrock, "Adaptive routing techniques for store-and-forward computer-communication networks," in *Proc. Int. Conf. Commun.*, 1971, pp. 39-1–39-8.

[5] M. Gerla, "Deterministic and adaptive routing policies in packet switched computer networks," in *Proc. 3rd Data Commun. Symp.*, Nov. 1973, pp. 23–28.

[6] I. Gitman, R. M. Van Slyke, and H. Frank, "On splitting random access broadcast communication channels," in *Proc. 7th Hawaii Int. Conf. System Sciences, Subconference on Computer Nets*, Jan. 1974, pp. 81–85.

[7] I. Gitman, "On the capacity of slotted ALOHA networks and some design problems," *IEEE Trans. Commun.*, pp. 305–317, Mar. 1975.

[8] R. E. Kahn, "The organization of computer resources into a packet radio network," in *Proc. Nat. Comput. Conf.*, May 1975, pp. 177–186.

[9] L. Kleinrock and S. S. Lam, "Packet switching in a slotted satellite channel," in *Proc. Nat. Comput. Conf.*, June 1973, pp. 603–710.

[10] L. Kleinrock and F. Tobagi, "Random access techniques for data transmission over packet switched radio channels," in *Proc. Nat. Comput. Conf.*, May 1975, pp. 187–201.

[11] H. Miyhara, T. Hasegawa, and Y. Teshigawara, "A comparative evaluation of switching methods in computer communication networks," in *Proc. Int. Conf. Commun.*, June 1975.

[12] Network Analysis Corporation, "The practical impact of recent computer advances on the analysis and design of large scale networks," 2nd semiannual tech. rep., Dec. 1973.

[13] —, "The practical impact of recent computer advances on the analysis and design of large scale networks," 3rd semiannual tech. rep., June 1974.

[14] Sessions on "Advances in packet radio communication," and "Packet radio—Future impact," in *Proc. Nat. Comput. Conf.*, May 1975, pp. 177–261.

[15] L. G. Roberts, "Dynamic allocation of satellite capacity through packet reservation," in *Proc. Nat. Comput. Conf.*, June 1973.

Packet Radio Network Routing Algorithms: A Survey

Jonathan J. Hahn
David M. Stolle

The increasing importance of PRnets in the local distribution of information over wide geographic areas

0163-6804/84/1100-0041 $01.00 © 1984 IEEE

Introduction

PACKET RADIO is a technology that has evolved from conventional networks connected by leased telephone lines using packet switching to transmit information. This technology yields an efficient way of using multiple-access radio channels to support communications among a potentially large number of mobile subscribers; it provides local distribution of information over a wide geographic area. In particular, packet radio networks (PRnets) lend themselves as attractive solutions for: 1) mobile resources such as terminals and computers that make broadcast methods necessary; 2) resources located in remote or hostile locations where the telephone system is not feasible, poorly developed, or uneconomical; and 3) where traffic characteristics of resources are of a bursty nature, for example, when there is a high ratio of peak bandwidth to average bandwidth. PRnets also provide a fairly good solution for local area networks (LAN's) in urban areas, such as the University of Hawaii's ALOHA system [1].

Packet radio networks also offer an advantage of increased bandwidth over conventional cable network systems. When shared among a large number of users which may frequently relocate, a single, high-capacity channel can be more efficient than a large number of fixed, low-capacity channels with mostly wasted capacity. The channel can be shared either by partitioning the channel into separate nonoverlapping frequency subbands or by scheduling each transmission in short nonoverlapping intervals. In the first case, known as frequency division multiple access (FDMA), each node has access to a dedicated portion of the channel at all times. In the second case, known as time division multiple access (TDMA), each node has dedicated access of the entire channel for only a portion of the time. However, the increase in bandwidth does not come without sacrifice. A channel in a *radio network cannot be explicitly associated with two* specific nodes. Packet transmissions may be received by all nodes within the range of the sending node. If node *i* wishes to send a packet to node *j* only, then it must add information to the packet header which instructs all other nodes to discard the packet. In addition, if a node receives a packet which contains an error, then the node may not know that it was the intended destination; thus, the destination node cannot request a new retransmission.

The routing and flow control strategies used to forward and control traffic are of utmost importance for packet network operation. These algorithms basically have three common objectives:

- reliability: to assure, with a high probability, that a message launched into the network will arrive at its destination;
- efficiency: to assure that messages will be delivered with a relatively small time delay; and
- low overhead: to assure that control traffic does not consume large amounts of channel capacity.

Without efficient routing and flow control, the same problems which affect conventional network systems can impede the processing of a radio network with a large number of users: a packet may circulate endlessly among the nodes, numerous copies of a packet may be circulating simultaneously, and, if the destination node is within another network, then several gateways may simultaneously receive a packet and introduce duplicates into the network. Routing algorithms in PRnets are typically concerned with an optimization that minimizes the length of a routing path (the number of hops), in hopes that this algorithm will also minimize the delay. Unfortunately, this is not necessarily true. Problems inherent in radio systems—such as frequent topological changes, quality of the transmission links, hidden terminal problems, and, particularly, the instability of packet routing information due to the mobile environment—often lead to a faulty delay minimization function.

Typically, a radio network contains two types of nodes. A node with origination, destination, and relay functions is called a repeater. A node with additional processing capabilities such as control, initialization, and accounting is called a station. Not all nodes must perform all functions in order to serve mobile terminals within a large geographical communications network. Repeaters with limited capabilities can provide area coverage, while a station can control capacity and reliability requirements among the several terminals within its range.

Unfortunately, the use of store-and-forward repeaters has some disadvantages. Repeaters generally use the same frequency for input as for output. Therefore, a repeater cannot start a retransmission until after it has completely received and stored a message. Furthermore, the repeater cannot receive any additional packets while retransmitting. Thus, when designing a network with fixed repeaters, both the order of packets to be transmitted as well as the topology of the repeaters must be considered.

Additionally, there are other classifications of networks. A network is a "single-hop" network if no relay functions are necessary at each of the network's nodes. Conversely, a network is a "multi-hop" network if packets can be relayed over several hops before reaching their destination. A network is a "stationless" network if it does not contain any stations. In this case, all the packet radio units (PRU's) are repeaters. In the most typical case, a "multiple station" network assumes processing control responsibilities at each of several separate stations.

Examples of routing algorithms which have been previously considered for packet radio networks include:

1) Broadcast routing [2,5], (second section)
2) Hierarchical routing [2], (third section)
3) Directed broadcast routing [2] and ARPANET-like routing [3,9], (fourth section)
4) Routing in a stationless network mode [3,7,8,10,11], (fifth section)
5) Multiple station routing [3,6], (sixth section).

This paper addresses these routing algorithms. Specific examples presented by Gitman et al., Khan et al., Kleinrock and Tobagi, and Perlman are cited.

Broadcast Routing

Gitman et al. [2] present a broadcast routing algorithm for single-station radio networks which essentially prevents packets from looping endlessly or cycling alternately between nodes. The algorithm is an especially useful technique to bypass the need for control of rapidly changing routes. This scheme may be used to flood an entire network or a large subset of the network. In general, this flooding method is a simple approach, which can be used to:

- change global parameters within all repeaters,
- map connectivity so that routing information can be updated and produce more efficient routing, and
- update routing information among mobile nodes which otherwise would be unfeasible.

The algorithm contains several mechanisms to efficiently route packets within a radio network which uses random transmissions and omnidirectional antennas. The routing scheme contains three main mechanisms which are described below: 1) a hop-by-hop acknowledgment scheme, 2) a maximum handover number to limit the number of hops traversed, and 3) storage within each repeater to hold a maximum of L packet identifiers. Also, a time parameter at each repeater may be used to discard any previously transmitted packets which return before a timeout has occurred. Thus, extra processing is not needed to search the L packet identifiers to determine if the packet has already visited this node. In addition, a variable transmission power mechanism can be used to increase the number of potential receivers and reduce the amount of required routing. However, if the transmission power is increased, then the amount of interference within the network will also increase.

The hop-by-hop acknowledgment is used to guarantee that a packet has been accepted by the next repeater. However, the acknowledgment does not need to be transmitted as a separate packet, since retransmissions by the receiving node are appended with an acknowledgment and will be received by the original sender. However, it should be noted that this free acknowledgment strategy does not work for the last hop. Thus, the destination node must send an explicit acknowledgment to its previous sender. This strategy assumes that all nodes can receive a transmission from a given node if they can send to that node. This method is more efficient than transmitting a separate acknowledgment packet to each neighbor, because each packet can acknowledge several neighbors simultaneously.

A handover number is used to avoid the problem of packets endlessly looping throughout the network. Initially, the handover number is set to M. Upon the reception of a packet, each repeater decrements the handover number by one; thus, the packet is assured of traversing no more than M transmissions. Unfortunately, if a station has only an approximate idea of the network topology, the packet may either never arrive or generate excessive duplicates, wasting bandwidth. However, as the handover number is allowed to increase, more alternate routes become feasible, making the network less vulnerable to repeater failures. Therefore, the choice of the initial value of the handover number is a critical design issue.

The final main broadcast mechanism is the storage of packet identifiers. Each repeater has a queue which stores a packet ID and time stamp of the most recently transmitted L packets. Whenever a packet is received, its identifier is compared with the identifiers stored within the queue. If a match occurs, then the packet is discarded (since it is a duplicate that has recently been transmitted). Each new packet is placed into the head of the queue and replaces the oldest identifying information. Consequently, only the most recent L identifiers are contained in the queue.

An improved broadcast routing algorithm is presented by Kleinrock and Tobagi [5]. This algorithm uses a numbering scheme which controls flooding using handover numbers.

The model assumes a uniformly spread topology of repeaters covering the entire region. Additionally, it is assumed that when a repeater transmits a packet, only those neighbors within its transmission range may receive the packet correctly.

The following numbering scheme allows a controlled flooding of the network. First, each repeater P_i is assigned a number N_i equal to the minimum number of hops between the repeater and the station (see Fig. 1). Second, when a repeater receives a new packet, the repeater assigns a handover number M such that $M >= N_i$. The repeater R_i then transmits the packet, decreasing M by one. Third, when another *repeater, R_j, receives a packet, it checks the handover* number M. If $M < N_j$, then the packet is either destroyed or ignored, since it is unable to reach its destination. If $M >= N_j$, then the packet is retransmitted, decreasing M by one. Once again, M limits the number of distinct repeaters that can handle a packet, specifically, only neighboring repeaters with a distance $M >= N_j$ from the point of origin. As M increases, so does the number of alternate routes available, decreasing network vulnerability.

Therefore, flooding represents a reasonable alternative when repeaters are highly mobile and network connectivity is unknown. The algorithm is both simple and reliable. There is a high probability that packets will reach their destination when the *route actually* exists and the handover number is large compared with the shortest path to the destination. The obvious disadvantage is the inefficient use of network resources, especially the large amount of bandwidth consumed. Also, the probability of duplicate packets generated at a gateway of another network may be high, unless prevented by gateway communications protocols.

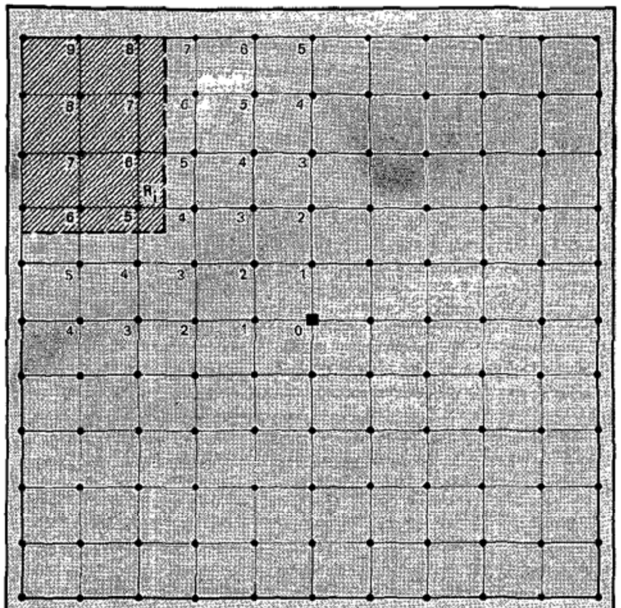


Fig. 1. **Numbers N_i assigned to repeaters R_i. Repeaters are designated by (●) and the station by (■).**

Hierarchical Routing Algorithm

The broadcast routing algorithm transmits data packets in all directions, resulting in inefficient use of the total bandwidth. In the hierarchical routing algorithm for single station networks, as presented by Gitman et al. [2], the repeaters are organized into a tree with the station at the root. This algorithm is particularly well suited for local collection and distribution of traffic primarily flowing from repeaters to stations and from stations to repeaters. If the traffic is to be routed between repeaters, it must pass through a station—creating a hierarchical routing requirement. Unlike the broadcast routing described earlier, this hierarchical routing requirement creates centralized control.

In single-station hierarchical routing, it is required that the central site be aware of the complete network topology. This information is acquired by having the central site send broadcast probe packets periodically. Each repeater responds to a probe by sending an answer packet. When a repeater forwards an answer packet due to another repeater, it appends its identification to the packet. From the returned answers, the central site can easily determine the shortest path to each repeater. Upon learning the new topology, the central site assigns routing information to repeaters. This information assigned by a station is called a label. The set of repeater labels forms a hierarchical tree structure of repeaters rooted at the station, as shown in Fig. 2. The labels may be changed during network operation when changes in network topology occur.

The information in each label includes: the routing address of the repeater, the minimum number of hops to the station, and the names of *all* repeaters on a shortest path to the station. Thus, the address of the next repeater along the transmission path is easily accessible.

A label consists of H fields, where H is the number of levels in the hierarchical tree. The label of a repeater R at level j (j-1 hops to the station) contains nonzero integers for

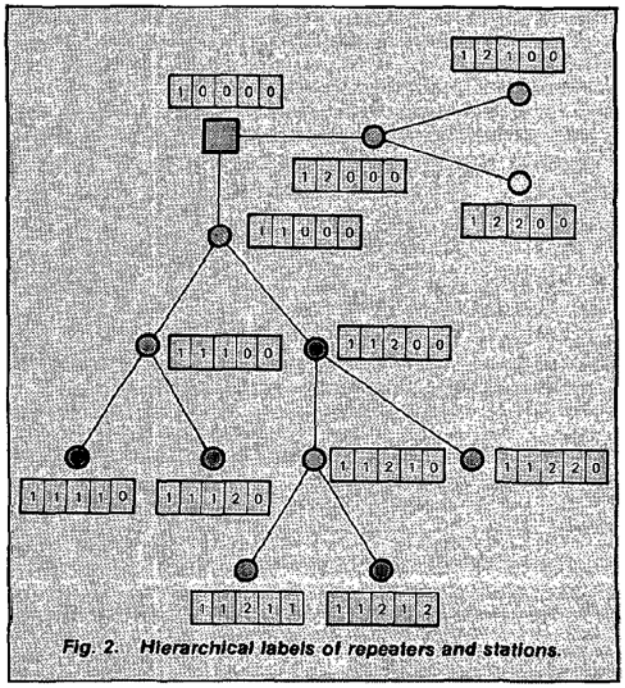


Fig. 2. **Hierarchical labels of repeaters and stations.**

the first *j* fields and zeroes in the remaining fields. The home of *R* is defined to be the repeater to which *R* addresses its packets when routing towards the station. For repeaters at level *j* which have the same home repeater, the labels differ only in the entry in field *j*. Thus, the label of the station has a nonzero entry in the first field and a zero in all other fields; the labels of repeaters at a distance of one hop to that station have the station's entry in the first field, nonzero unique entries in the second field, and zero in all other fields. Figure 2 shows an example of a labeled set of repeaters.

Once all the repeaters have been labeled, each data packet sent by the central site can contain the label of the repeater to which the station has been assigned. The packet must also contain a pointer indicating the current field in the label. When a packet arrives at a repeater, the packet label's current field and the repeater label's corresponding field are checked for a match. If they match, then the repeater either decrements or increments the current field indicator, depending on whether the packet is being directed to or from the station, and forwards the packet. If not, the repeater just discards the packet.

If a repeater along the required path has failed, all is not lost. When a repeater exceeds its allowed number of retransmissions without receiving an acknowledgment, it can set a bit in the header instructing all repeaters to adopt the flooding algorithm for the packet. This begins an alternate routing procedure and a failed or temporarily busy repeater may be bypassed. It could also announce the failure to the central site, requesting the central site to conduct another probe and relabel the tree.

Another alternate routing method when blocking is encountered is simply to turn up its transmission power and hope to skip over the failed repeater. However, there are some limitations to this method:

- Duplicate copies of the packet may be generated.
- An increase in power results in interference with a large number of repeaters.
- Although the correct repeater may receive the packet, the repeater's transmission power may be too weak to send the acknowledgment back to the sender.

Unlike the broadcast algorithm, the hierarchical routing algorithm allows shortest path routing between repeaters and stations when the labels are properly assigned. This results in less utilization of nodal processing capacity and better use of channel capacity. Thus, average delay experienced for the hierarchical algorithm will be lower than that for the broadcast algorithm, as only repeaters on the path transmit the packet. This yields smaller interference probability, fewer transmissions per hop before success, and consequently a smaller delay per hop. There is also a high probability that the packets will reach their destinations. Since routing is directional, packets will arrive at only one station. One drawback is that the algorithm requires some maintenance of an updated set of labels, but this becomes a limitation only in a network with highly mobile nodes.

Directed Broadcast Routing

The hierarchical routing algorithm is most useful when the radio network is used for local collection and distribution of traffic or when the number of nodes is very large. Now suppose that: 1) traffic flows are equally likely between any pair of nodes; 2) the traffic origination device knows (or can obtain) the ID of the destination repeater; and 3) that accounting information for each packet or message is not needed. With these assumptions, a distributed network architecture is appropriate. Hierarchical routing can still be used, but distributed routing—whereby a packet can be routed directly to the destination repeater—may be more efficient.

In a distributed radio network, stations act mainly as observers and controllers performing:

- network initialization, mapping network topology, determining labels, and initializing repeaters;
- network monitoring, observing connectivity and changing repeater labels;
- global flow control functions and controlling operating parameters of repeaters;
- gateway functions; and
- supervisory functions.

As presented by Gitman et al. [2], in a directed broadcast, a repeater only forwards a packet if it is closer to the destination than the last repeater that forwarded the packet. Each repeater is assumed to know its distance in hops from every other repeater. This information can be acquired by having each repeater broadcast its distance table periodically to the station. Each data packet contains the identification of the destination and the sender's distance from that destination. There is a new sender at each hop, and hence a new distance. When a packet arrives at a repeater, a check is required to determine if the repeater is closer to the destination than the sender. If so, the packet is heading in the correct direction and is forwarded. If not, the packet is headed the wrong way and is discarded. When encountering blocking, a repeater may increase its transmission power. However, this may result in some difficulties in obtaining an acknowledgment, as discussed in the previous section. As a last option, the repeater may send a control packet to a station to resolve the difficulty.

Another type of directed broadcast routing algorithm is the ARPA packet radio routing algorithm [3,9]. Unlike the scheme described above, this algorithm does not rely on a central station to distribute global routing information to each repeater. In a distributed manner, repeaters maintain their own connectivity and delay matrix. Periodically, each repeater announces its existence by transmitting a distance vector and other status information to its neighbors. The distance vector contains an estimate of the minimum delay to every repeater in the network. When a data packet is transmitted, it contains the identification of the destination and the identification of the next repeater en route. All other neighbors which are not the designated receiver discard the packet. Therefore, the packet is only transmitted in the direction of the destination. Unfortunately, distributed updating of routing information generates an excessive quantity of control traffic. Thus, network efficiency will be greatly reduced, especially when the routing tables must be frequently updated due to highly mobile stations and repeaters. Another disadvantage of the scheme is that the routing tables grow without bound as the number of nodes within the network becomes large.

Both of the directed broadcast algorithms described above enable shortest-path direct routing (that is, not via a station) to any repeater or station in the radio network with a high probability of reaching the destination node. These schemes are also more robust than the hierarchical routing because they concurrently try all alternate paths whose lengths are equal to the minimum length. However, these procedures consume more bandwidth. Another point in their favor is their ability to adapt to mobile repeaters. On the other hand, these algorithms require a need to maintain an updated set of labels or distance vectors, and this becomes the main limitation in a network of numerous, highly mobile nodes. If the application implies a distributed architecture and the repeaters and stations are mobile, the task of updating distance vectors should be distributed. This task will require more repeater capabilities.

Stationless Routing

Khan et al. [3] and Perlman [7,8] introduce an approach for effective routing in a "stationless" network environment. When no stations are present in the network, the main difference in operation is that each packet radio (PR) must determine network connectivity on its own. Typically, each radio initially relies on a broadcast routing scheme to communicate, even though this procedure is known to be somewhat inefficient. The radios must determine a satisfactory point-to-point route, even if the route remains acceptable for only a short time. Repeaters and terminals prepare packet transmissions in three phases:

- A route finding packet (RFP) is broadcast by the source PR to create several path possibilities.
- A route set-up packet (RSP) is transmitted by the destination PR back to the source PR, so that the intermediate radios may store appropriate routing information.
- The normal data packets are transmitted by the source PR using a small packet header to indicate the next few hops to be taken along the route.

An RFP is transmitted when the source PR intends to transmit to an unknown destination. The packet contains the source ID, the destination ID, a unique identifier, and a list of PR ID's. During its route traversal, any radio which hears the broadcast adds its own identifier onto the ID list, increments the hop counter, stores some identifying route and delay information in both buffer memory and the packet header, and rebroadcasts the packet. In order that the network does not become congested, several measures are taken. Each PR discards any duplicates of a packet previously received. Additionally, if the hop count exceeds a maximum value, then that packet is discarded. A packet is also discarded when a radio has already sent a packet which contained a smaller delay. A problem arises when an RFP never reaches the destination. The problem is primarily due to the inherent instability of radio networks, which occurs since no acknowledgments for RFP's are transmitted. Consequently, the source PR may find it necessary to attempt retransmissions after a timeout has been generated.

If the RFP reaches the destination, then a successful route possibility has been identified. Several RFP's should arrive, so many possible routes, each carrying its own delay estimate, will be suggested. The destination PR waits some amount of time from the first RFP received, chooses the route with the minimum delay, and stores this route as the optimum. Finally, a route set-up packet is initiated.

A route set-up packet is sent from the destination to the original source to set up the optimal route selected above. The packet contains the source ID, the destination ID, a portion of or the entire route sequence, and a total hop count. The packet traverses the route in reverse, setting up intermediate routing information at each PR until the packet reaches the source. It may be necessary for an alternate route to be taken, if a radio fails. When the packet is traversing an alternate route, a radio which hears the packet and has a good neighbor that is further down the route may continue the packet along its path. When the source receives the RSP and the route has been initialized, the data packet may be transmitted.

Since the created path may involve several radios, not all routing information is stored in the packet header. Some routing information is stored in the intermediate PR's along the path. In the data packet transmission phase, acknowledgments are used on a hop-by-hop basis. Thus, a radio may send a failure notice to the source—if a radio fails to receive an acknowledgment from the next PR on the route, or if the radio does not contain enough buffer memory to store the route information. If an acknowledgment is not received and a failure notice is produced, then an alternate route must be found by transmitting to neighboring radios. It is hoped that the packet can be directed back onto the route as planned. However, if a failure results due to the lack of buffer memory, then a new route must be found, causing the entire procedure to be redone. Unfortunately, problems may become even more compounded, since there are circumstances when a failure notification (route loss packet) may never be received by the source.

The final step involves transmitting the data packet. A small routing header is used to specify the next few hops to be taken. Each radio overwrites the route header information from the route information stored within its own memory. Thus, new route information for upcoming hops is given at each radio. So, if the next PR in the route sequence has failed, then an alternate route may be taken. To prevent looping, there is a limit on the number of alternate routes which may be taken consecutively.

Therefore, the routing algorithm for a stationless network is both complicated and redundant. Channel usage and contention are high, since an RFP must be broadcast initially. Route selection is cumbersome; even worse, the selection process may not include the path containing the minimum delay. Finally, the network does not maintain a congestion control mechanism, so long delays may exist when transmitting packets. Also, if the network topology is rapidly changing, an acceptable path may never be found. In such cases, a flooding algorithm may be adopted. However, even if the network topology is not known, the algorithm does provide a route.

Multiple Station Routing

In the normal case of multiple-station operation, the responsibility of network control is distributed among all stations. Specifically, the stations provide control for the mobile terminals within their range. If one station fails, the others are to temporarily assume its functions with little or no degradation in system performance. Multistation operation is therefore more efficient and robust than single-station operation, especially when the number of terminals within the network becomes large or the station fails.

Each PR broadcasts a radio-on packet (ROP), which contains status and identification information. A set of neighboring radios within range will hear this ROP, noting the signal strength of the transmission. Each radio holds a cumulative count of the number of packets received from the other radios. Thus, each radio can determine the set of radios with which it can communicate reliably. Once this information is established, each radio periodically sends summary ROP's containing labeling and neighbor table information to reachable stations.

Upon hearing an ROP from a radio, a station will label that radio. The labeling process consists of a station determining and then supplying a radio with a route, which is an ordered set of radios called selectors, to that station. Once labeled, the radio periodically transmits summary ROP's, indicating those stations which are in direct range of the radio. Having been supplied with this information, each radio has labeling slots for storing routes to several stations. The station must relabel the radio within a given time or the labeling slot entry will expire. A slot whose entry has expired is not erased by the radio, but it may be overwritten by another station if no other labeling slots are available.

Kahn et al. [3] present an algorithm for such multiple-station operation. Each station is assumed to know which radios it has labeled, but must communicate with other stations to learn the location of other radios not under its control. This exchange of label information provides the destination station with a set of selectors for a point-to-point route from the originating radio or station to the destination. If both users are highly mobile, a point-to-point route will be established between the two end stations. These stations will then individually handle the final distribution. On the other hand, if both users are fixed, the destination station will choose the last few selectors to obtain an optimal route directly between the two end users. Lastly, if one user is fixed while the other is mobile, the resulting point-to-point route will be between the fixed user and the remote station. This station will then handle the local distribution for the mobile user.

For the route-selection process, a radio generates a packet for a destination outside the control of its stations and routes the packet to an appropriate local station. This station converts the packet into several distinct RFP's. These packets are sent to each neighboring station through some repeater jointly labeled by both stations. Each RFP includes the station ID and a list of selectors from the source PR to the jointly labeled repeater. When the packet is received by a neighboring station, it determines if the destination radio is under its control. If not, the station modifies the received RFP and converts the packet into several distinct packets by adding its own ID and a list of selectors from the original jointly labeled repeater to another repeater jointly labeled by a station not previously visited by the packet. If the packet arrives at a station which has previously handled the same request, then the packet will be discarded. The destination station then transmits a complete list of selectors to the destination radio, which initiates the route set-up procedure. To set up the route, the destination radio sends a route set-up packet back to the source radio. This packet is used to update the routing tables at each intermediate node, by providing the complete list of selectors obtained from the route finding procedure. Figure 3 illustrates this process for a source-destination pair *A,B.* The RFP is shown as a series of dotted lines connecting stations, and the point-to-point route is shown as a series of solid lines connecting repeaters.

If a radio fails while a packet is en route, disrupting a point-to-point route, then local alternate routing around the failed radio will take place, if possible. The station in control of the failed radio will attempt to take a corrective measure using only the routing selectors contained in the en-route packet and the destination ID for transit traffic. A station's transit traffic consists of those packets whose destination has not been labeled by that station. If the corrective measure fails, an error message will be transmitted to the source and the original route set-up process will be reinitiated. In this process, packets may be discarded, so the end-to-end protocol must be prepared for recovery.

Thus, multistation routing is more complex when compared to routing used by single-station networks. Yet, the algorithm is designed for supporting several geographically distributed users, possibly thousands or more. The scheme is more efficient than stationless routing, since each station locally controls several users. In the event of a station failure, control is decentralized, since other stations may resume the operation of the failed station.

Conclusion

The packet radio has demonstrated that it is a possible technology for use in both fixed and mobile computer communications. A packet radio still remains too expensive for most commercial use; however, recent advances in integrated circuit technology and signal processing techniques are expected to reduce the cost. Therefore, packet radio will become increasingly important in the local distribution of information, especially when the source and destination are constantly changing. Also, packet radio offers itself as a significant alternative when designing a local area network.

In the case of single-station network operation, the three classes of algorithms which were presented contain both advantages and disadvantages. The broadcast algorithm is the most inefficient, yet no initialization is necessary and the

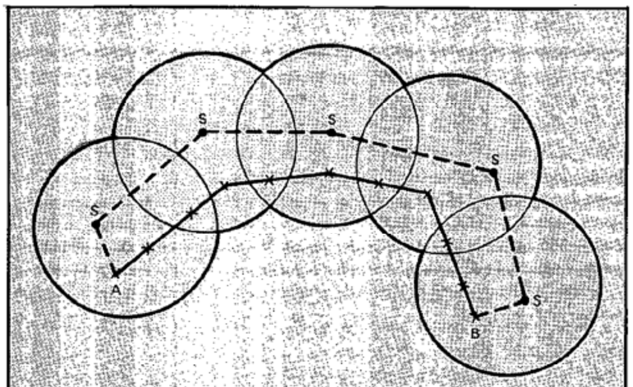


Fig. 3. Route finding and route set-up with multiple-station operation. Stations are designated by (●) and repeaters by (X).

amount of control information within the packet header is quite low. This flooding technique is especially useful for networks with rapidly changing connectivities. The optimum choice between the directed broadcast algorithm and the hierarchical routing scheme is not immediately apparent. The directed broadcast algorithm is more reliable, since more than one receiver is able to accept a packet for switching. Also, the hop-by-hop acknowledgment scheme will reduce the number of possible paths and duplicate packets, especially near the destination. However, on a path between a station and a repeater, this algorithm does utilize more resources. On the other hand, the hierarchical algorithm is more adaptable to changes in network connectivity, with lower overhead for updating repeater labels. For this algorithm, when a new repeater is added, only the new repeater need be initialized, while the directed broadcast algorithm must change all repeaters' labels. Furthermore, since all packets are routed through a central station in the hierarchical method, system monitoring and control are handled more easily. However, both methods are vulnerable to the loss of their central station.

There are four major differences in the functions of a network which contains stations and one which does not. First, a station constantly collects network connectivity information, so that it may supply a route to the source PR immediately. Thus, it is not necessary to broadcast an RFP. Second, a station can compare all possible routes when making a choice, not just the ones which traversed the path successfully in a single attempt. Since there are more possibilities from which to select the route, a more efficient route will be selected. Third, a station can detect changes in network connectivity and adjust routing appropriately. Therefore, the station does not interrupt ongoing communications among radios. Finally, a station can perform global congestion control by changing parameters within the individual radios of a surrounding area. In contrast, the routing within a stationless network is inefficient and involves more of the network resources. However, there may be instances when the radios are made operational before the stations or when the stations no longer exist. In this case, a stationless mode of operation would be very useful.

References

[1] N. Abramson, "The ALOHA system—another alternative for computer communications," *AFIPS Conf. Proc., Fall Joint Comput. Conf.*, vol. 37, pp. 695–702, 1970.

[2] Gitmán et al., "Routing in packet switched broadcast radio networks," *IEEE Trans. Commun.*, Aug. 1976.

[3] R. E. Khan et al., "Advances in packet radio technology," *Proc. IEEE*, vol. 66, no. 11, Nov. 1978.

[4] L. Kleinrock and A. Belghith, "Distributed Routing Scheme in Stationless Multi-Hop Packet Radio Networks: 'Mobility Handling'," UCLA, June 1982.

[5] L. Kleinrock and F. A. Tobagi, "Routing with handover numbers," PRTN 11 Dec. 1973.

[6] W. I. MacGregor, "Flooding in multistation routing," PRTN 303, June 1981.

[7] R. Perlman, "Stationless compatible PRnet routing," PRTN 256, June 1978.

[8] R. Perlman, "Remaining issues in stationless compatible routing," PRTN 258, Sept. 1978.

[9] A. S. Tanenbaum, *Computer Networks*, Englewood Cliffs, NJ: Prentice-Hall, Inc., 1981.

[10] J. Westcott, "Station to stationless network interface," PRTN 300, June 1981.

[11] J. Westcott and J. Jubin, "A distributed routing design for a broadcasted environment," *Proc. MILCOM '82*, pp. 10.4-1–10.4-5 Oct. 1982.

Jonathan J. Hahn is currently working as a member of the technical staff in the Operating Systems and Networking Department at TRW, Inc., Redondo Beach, CA. He received his B.S. in Information and Computer Science from the University of California, Irvine, in 1981 and his M.S. in Computer Science from the University of California, Los Angeles, in 1984. He is a Member of the IEEE Computer Society.

David M. Stolle received his B.S. in Computer Science and Mathematics from Vanderbilt University, Nashville, TN, in 1981 and his M.S. in Computer Science from the University of California, Los Angeles, in 1984. Since 1981, he has been a member of the technical staff at TRW Inc., Redondo Beach, CA, designing and developing system software. In June of 1984, he became a member of the technical staff in the Strategic & Information Systems Division at Logicon Inc. He is a Member of the IEEE Communications Society. ■

Issues in Packet Radio Network Design

BARRY M. LEINER, SENIOR MEMBER, IEEE, DONALD L. NIELSON, MEMBER, IEEE, AND FOUAD A. TOBAGI, FELLOW, IEEE

Invited Paper

There are many design choices that must be made in the development of a packet radio network. There is usually no single correct choice, and the decisions are dependent on the environment that the network must work in, the requirements for performance and other functionalities, and the cost and other limitations. In addition, as new hardware and software technologies become available, the parameters governing the decisions change and often result in different selections.

This paper outlines a number of design issues and choices available. The intent is to provide an overview of the design decisions that must be made so as to provide a context for the decisions made in a number of existing and developing packet radio networks. It is hoped that this will allow future designs to take advantage of both the wealth of experience available as well as new technologies. Three areas of design decisions are identified. The first area deals with the physical aspects of the network and concentrates on the radio connectivity and channel sharing. The second area deals with the automated management of the network and concentrates on issues such as link management and routing. The third area deals with the interface of the network to the users and some practical aspects of operating and maintaining a network.

I. Introduction

Packet radio networks represent the extension of packet switching technology into the environment of broadcast radio. They are intended to provide data communications to users located over a broad geographic region, where direct radio or wire connection between the source and destination users is not practical.

Packet radio networks have been and are being designed to operate in a number of environments using a number of different technologies [1]. There are packet radio networks making use of ground mobile radio (narrow-band 16 kbits/s [2] and more wide-band 400 kbits/s [3], [4]), amateur radio [5], HF for use in Navy applications [6], and satellites [7]. Yet all of these networks share some common characteristics. They are all based on the notion of packet switching applied to (usually broadcast) radio usually sharing a single channel. They are intended to handle mobile users, although some of the amateur and commercial applications de-emphasize this capability. They are for the most part based on a store-and-forward operation, although the simpler satellite networks involving use of a single satellite do not include store and forward operation.

Fig. 1 shows a typical packet radio network structure [8]. A packet radio unit consists of a radio, antenna, and digital controller. The radio provides connectivity to a number of neighboring radios, but typically is not in direct connectivity with all radios in the network. Thus the controller needs to provide for store-and-forward operation, relaying packets to accomplish connectivity between the originating and destination users.

There are a number of common issues involved in the design of these networks. These include efficient methods for sharing the common radio channel, methods for de-

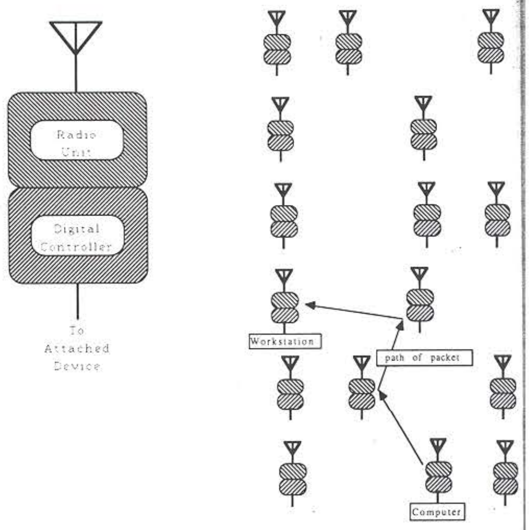


Fig. 1. Packet radio network structure.

Manuscript received June 20, 1986; revised August 18, 1986. The views presented in this paper are those of the authors and do not necessarily represent those of their organizations or the sponsoring organizations.

B. M. Leiner is with RIACS, NASA Ames Research Center, Moffett Field, CA 94035, USA.

D. L. Nielson is with SRI International, Menlo Park, CA 94025, USA.

F. A. Tobagi is with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA.

0018-9219/87/0100-0006$01.00 © 1987 IEEE

termining connectivity and using that connectivity to route data through the network, methods for achieving reliable communciations in a typically noisy radio environment, and methods for managing and controlling the distributed network. Thus there are many design choices that must be made in the development of a packet radio network. There is usually no single correct choice, and the decisions are dependent on the environment that the network must work in, the requirements for performance and other functionalities, and the cost and other limitations. In addition, as new hardware and software technologies become available, the parameters governing the decisions change and often result in different selections.

This paper outlines a number of design issues and the various choices available. The intent is to provide an overview of the design decisions that must be made and to do so in the context of a number of existing and developing packet radio networks. It is hoped that this will allow future designs to take advantage of both the wealth of experience available as well as new technologies. Furthermore, as the rest of this Special Issue will provide details on a number of packet radio systems, this paper should provide a context for comparison of the various approaches taken in those systems.

Three areas of design decisions are identified. The first area deals with the physical aspects of the network and concentrates on the radio connectivity and channel sharing. The second area deals with the automated management of the network and concentrates on issues such as link management and routing. The third area deals with the interface of the network to the users and some practical aspects of operating and maintaining a network. The various issues are highlighted throughout the paper by showing them in raised or capitalized text. A compilation of these issues should help guide the design of a packet radio network.

Just as with any packet communication system, the functions to be performed by a packet radio network may be organized into a linear hierarchical structure, as defined in ISO's OSI Reference Model [9]. This structure consists of a layered architecture comprising a number of independent layers. This allows the discussion of design issues underlying a packet communication network to be done by focusing on one layer at a time. While there will always be coupling between the various layers in terms of design issues, requirements, etc., in the case of packet radio (as will be clear from the discussion below), the various layers are highly interdependent. Therefore, the task is not complete until a cross examination of design tradeoffs at all layers is performed. Nevertheless, to achieve an orderly presentation, the issues are presented one layer at a time.

II. Physical and Data Link Layers

In this section, we focus on the first two layers of the ISO model. Issues to be addressed include physical connectivity, bandwidth–time–space management, channel access, and data link control.

A. Physical Connectivity

The physical layer in packet radio networks establishes digital link connectivity among nodes in the network so that information (data) paths may be established from traffic sources to their destinations. Link connectivity from a node *A* to another *B* refers to *B*'s ability to correctly receive information transmitted by A at a specified minimum rate. Link connectivity clearly depends on radio propagation parameters, such as the radio frequency, the distance between nodes, the type of terrain, and the transmit power. Connectivity depends in addition on the data rate requirement, the channel RF bandwidth, and the data encoding and modulation schemes. Thus the design problem at this level can be formulated as follows. GIVEN THE TYPE OF TERRAIN IN WHICH THE NETWORK IS TO OPERATE, THE USERS' LOCATIONS, AND THEIR REQUIREMENTS IN TERMS OF TRAFFIC, MOBILITY, ANTIJAMMING CAPABILITIES, ETC., SELECT

A. THE RADIO FREQUENCY AND RF BANDWIDTH,
B. THE SIGNALING, ENCODING, AND MODULATION SCHEMES,
C. THE NETWORK TOPOLOGY.

Network topology refers to the radio nodes constituting the network, their roles (user interface or repeater), their density, their locations (fixed, mobile), the antenna design associated with each (its height, directionality, etc.), and their transmit powers.

The connectivity resulting from the design decisions may be represented as a graph in which vertices represent nodes, and edges represent link connectivity. Due to changes in the environment, nodal mobility, and other effects, variations in digital link connectivity may result, rendering the graph representing the network topology a probabilistic one.

Design decisions at the physical layer interact with those at higher layers. For example, propagation is typically better at lower frequencies (longer distances can be achieved with less sensitivity to terrain). On the other hand, data rates are lower at lower frequencies, and therefore, the network's ability to cope with mobility may be reduced. Thus if the user data requirements can be satisfied with relatively low data rates, a low frequency can be used, thereby possibly achieving full link connectivity and simplifying the network design. On the other hand, higher data rates would require a higher frequency/bandwidth and probably line-of-sight (LOS) propagation. This would result, though, in more capability to support the overhead of network algorithms to handle changing connectivity resulting from mobility (coupled with LOS propagation.)

Nevertheless, by initially ignoring effects due to higher level functions, an approximation to the design can be achieved, which then can get refined as the higher level design issues are resolved. For example, it is possible to estimate a range for the required link data rates given the users' traffic and delay requirements, independently of higher level protocols, and proceed with the design based on these estimates. A more precise determination of the required rates would then have to be done after an understanding of higher level issues has been acquired.

1) Choice of a Frequency Band: The first issue to be addressed is WHAT FREQUENCY BAND TO OPERATE IN? The tradeoffs between higher frequency/bandwidth LOS operation and lower frequency/bandwidth operation over extended distances, while present in the design of any radio system, have special implications in a packet radio network.

Not only does this tradeoff determine the degree of connectivity between the user nodes, but the choice of frequency may require additional functions to be supported.

For example, if a frequency with good propagation characteristics is chosen, sufficient connectivity may be achieved just using the radios at the user nodes. Higher frequencies, because of the limited propagation distances at LOS frequencies, may require "repeater" nodes be installed to achieve full network-wide connectivity. (A repeater node is a packet radio unit with no user directly connected.)

Another consideration in the choice of frequency is the available bandwidth. This choice is partially determined by the user data requirement as well as other factors such as the need for spread spectrum. However, the dynamics of the network topology, and therefore the required amount of control traffic, also help determine the required data bandwidth and therefore the minimum frequency band. Networks where the connectivity is changing slowly (e.g., HF networks in a Navy environment [6] or fixed-site networks such as the amateur packet radio networks [5]) can afford to have limited control traffic. Ground mobile networks [3], [4], due to their rapidly changing connectivity, require considerably greater control traffic and therefore increased data bandwidth.

2) Propagation and Interference Considerations: In addition to frequency band, link connectivity also depends on other factors, including terrain, distance between nodes, transmitter power, antenna height and directionality, etc. Given the terrain, antenna parameters, and node locations, link connectivity is achieved by appropriate selection of a data-encoding scheme and transmitter power. When the radio channel is being shared by many nodes, multiuser interference and the near-far problem are introduced. Because these issues pertain more specifically to the problems of bandwidth-time-space allocation and channel access, discussion is deferred until the next section.

B. Bandwidth-Time-Space Management

Once a frequency band and RF bandwidth have been selected, the question remains as to HOW TO ALLOCATE THE BANDWIDTH IN TIME AND SPACE TO THE NODES IN THE NETWORK. Four techniques are available, all of which may coexist: frequency division, time division, code division, and spatial reutilization of the bandwidth resources. Frequency division refers to the partitioning of the bandwidth into separate radio channels, orthogonal in the frequency domain. Time division refers to the allocation of a given radio channel to different users at different times. Code division refers to the provision of orthogonal spread spectrum codes to different (radio channel) users, so that these may use the same channel simultaneously without interference. Finally, spatial reutilization of the bandwidth-time resource refers to the simultaneous use of a given portion of the channel bandwidth in different localities without causing interference. A few examples are presented here to illustrate some design considerations and tradeoffs.

The use of frequency division in order to provide channels for use by individual nodes (or pairs of nodes) may be adequate if the period of use for channels is long, and the channel utilization high; that is, if the users' demand is nonbursty and predictable. Otherwise, overall bandwidth efficiency will be low. Furthermore, such a mode requires management procedures to dynamically allocate these channels. From another, perhaps more important, point of view, frequency division may be useful in the provision of several channels, each of which is used for a different functionality, but is shared in the time or code domains by many users [6]. This constitutes a natural way of providing a hierarchical network. (In the HF band, for example, radio distance varies significantly with the channel frequency. A channel centered around a frequency at the low end of the band could be used for overall network connectivity and control purposes, while other channels operating at higher frequencies are used for communication among neighbors.) But in general, it eases network deployment and resource allocation and management to have all users tuned to the same channel frequency, and employ time and code divisions [2], [3]. The specific means by which a radio channel is shared in the time and code domains are discussed below under channel access and capture modes. Suffice it to say at this point that such schemes are devised so as to achieve efficient bandwidth utilization. Furthermore, if nodes employ omnidirectional antennas, then local broadcast communication results, and mobile users are more easily supported without the need for complex network management procedures.

Spatial reutilization of the bandwidth-time resource achieves a higher overall utilization by allowing multiple transmissions to take place at the same time (and with the same code) in different geographic areas of the network. Such reutilization may either be the result of RF propagation characteristics, or intentionally designed for. Using the UHF band for ground mobile operation, for example, the radio range is inherently short, rendering spatial reuse a natural outcome. In this and other cases, reutilization may also be achieved by using directional antennas, or low transmission power.

It is not clear when to intentionally enforce spatial reuse. Consider, for example, the problem of using transmission power to control the range of nodes in a wide geographic area with a relatively dense population of nodes. If source-destination pairs are distant, then high transmit power leads to a smaller number of transmission hops, but a higher degree of interference with other nodes results, which in turn limit the network's throughput; a low transmit power leads to less interference, and thus a higher degree of spatial reuse of the bandwidth; but it then requires a larger number of transmissions to reach the destination, thus increasing channel load for the same user traffic, and thus perhaps again limiting the network's throughput [10].

C. Channel Access

Two mechanisms to share a single radio channel are time division and code division (in the case of spread-spectrum transmissions.) WHAT CHANNEL ACCESS PROTOCOL AND CODE ASSIGNMENT ALGORITHM SHOULD BE USED?

In sharing a channel, there is a need to deal with conflicts which result from contention. This is achieved either by *a priori* fixed assignment of the channel resources (time and codes) to different users so as to prevent contention, or by providing some dynamic channel access algorithm which defines when users are permitted to transmit based on channel conditions and traffic demands. Although such algorithms may not prevent conflicts from occuring, the objective is to maximize the overall network throughput. As channel access protocols and their performance are closely related to channel signaling methods and the capture effects that result, we will discuss them for narrow-band systems and spread-spectrum systems separately.

1) Narrow-Band Systems: Aside from the limited effects of power and FM capture, the overlap of two or more packets at some receiver in narrow-band systems results in the destruction of all. We say in this case that the system operates under a *zero-capture* mode.

At first glance, the solution appears to be one which guarantees orthogonality in the time domain. Fixed TDMA, whereby time slots are permanently assigned to nodes (or pairs of nodes), suffers from the same limitation in efficiency as does FDMA when nodes' traffic is bursty. Dynamic allocation of time slots to match traffic requirement is more efficient, but then requires scheduling algorithms which tend to become complex in distributed environments [6].

The time sharing of a channel by users can also be achieved via random-access schemes [10]–[12]. The decision as to whether to transmit or not is entirely left to the nodes, and thus collisions may occur. If a transmission collides with others, then it is repeated at some later time.

Random-access protocols are of basically two types, the ALOHA type and the carrier sense type. In the former, no knowledge of current activity in the network is required, and nodes act completely independently. The ALOHA scheme, which allows a node to transmit any time it wishes, is one such scheme [13]. In the carrier sense type, some knowledge of transmission activity in the network is acquired and used in the decision. For practically realizable protocols, the rules embodied in such protocols are constrained to be in terms of information that can be acquired locally at the node; typically, this consists of the transmission activity of neighboring nodes, which is acquired mostly via carrier sensing. Carrier Sense Multiple Access (CSMA), in which a node may transmit only if it does not sense any carrier (which otherwise would be due to neighbors transmitting), is one example [14].

While it may appear that CSMA would greatly outperform ALOHA, this is not always true and depends to a large extent on the network topology and traffic. In a fully connected network, in which the propagation delay is a small fraction of the transmission time of a packet, CSMA is significantly superior to ALOHA. But in a general multihop topology, the existence of hidden nodes (i.e., nodes within range of the intended destination but not of the transmitter) can drastically degrade the performance of CSMA [12]. Finally, we note that the implementation of CSMA also requires special hardware, and the ability to switch rapidly from the receive mode to the transmit mode in order to keep the efficiency of the scheme high.

The problem of collisions caused by hidden nodes can be alleviated by the use of a busy tone which is transmitted by a node on a separate channel to indicate that it is currently receiving a packet [15]. This activity-signaling channel requires additional bandwidth and hardware resources which increase the cost of the radios. Furthermore, deciding which node should transmit the busy tone and under what condition, and the possibility of blocking transmissions which would otherwise have succeeded without interfering with ongoing ones, complicate matters further.

The above mentioned schemes, and variants thereof which may prove appropriate for particular situations, offer tradeoffs and performance results that are not simple to assess. Performance analysis via mathematical modeling and simulation has been carried out for some schemes to some extent, leading to results which can be helpful in understanding their behavior [10], [12].

2) Spread-Spectrum Systems: SHOULD SPREAD-SPECTRUM CODING BE USED? The selection of a signaling method may be based on considerations other than digital connectivity or jamming. For example, spread spectrum may be used to combat multipath, or FH spread spectrum may be used to overcome the near–far problem. On the other hand, as will be clear from the discussion below, there may be advantages in using spread spectrum specifically for sharing a channel, because of the resulting reduction in multiuser interference. But in general, given that spread-spectrum coding requires wider bandwidth, it is not clear whether overall bandwidth efficiency is improved or not [16].

The main features which distinguish spread-spectrum systems from narrow-band systems are *code-division* and *time-capture*. Code-division refers to the fact that transmissions with orthogonal spreading codes may overlap in time with little or no effect on each other. Time-capture refers to the ability of a receiver to successfully receive a packet with a given code despite the presence of other time-overlapping transmissions with the same code. To simplify the discussion, no distinction is made between direct sequence pseudo-noise (PN) modulation and frequency hopping (FH) as the means for achieving spread spectrum, and multipath is ignored. In either case, it is assumed that the transmission of packets is asynchronous, and hence there is a need to precede the transmission of the packet by that of a preamble, which receivers use to acquire bit and packet synchronization. Furthermore, depending on the need, some form of data encoding for error correction may be performed to recover from erased or incorrectly received symbols.

Once the form of spread-spectrum modulation is selected, the method for ASSIGNMENT OF CODES USED IN THE PREAMBLE AND DATA PORTIONS OF THE PACKET must be selected. (The data portion includes control headers for network layers and above, as well as user data.) Two basic alternatives exist for the preamble codes. The preamble may consist of a known code with strong autocorrelation properties, which is used throughout the network and which idle receivers constantly search for. In this case, it is reasonable to assume that the overlap at some receiver of preambles belonging to different packets would cause errors in the processing of all preambles, and no packet is then acquired. Otherwise, the preamble is correctly processed (assuming that the background noise level is not too high), and the packet is locked onto. This case is refered to as the *space-homogeneous preamble code assignment*.

An alternative is to use preamble codes which are specific to intended receivers, refered to as the *receiver-directed preamble code assignment*. This alternative results in reduced preamble interference, as fewer packets would share the same code. However, information regarding the assignment of codes to receivers must be disseminated throughout the network. Furthermore, as we will see later, broadcast reception (the ability of all neighbors of the transmitting unit to hear the transmissions) can be quite helpful in dissemination of information for routing and network management.

Codes must also be assigned to the data portion of the packet. One desirable characteristic for the code assignment is that, once a packet is locked onto by a receiver, other overlapping packets do not interfere with its correct reception. Here too, there are several alternatives. In the *space-and-bit-homogeneous code assignment*, all data bits are encoded with the same code. In this case, an overlapping packet would not interfere with a packet locked onto as long as its autocorrelation peaks do not coincide with those of the earlier packet; i.e., unless the bit periods of the overlapping packets are within a few chip times of each other, causing the correlation peaks to overlap. As with preamble code assignment, this kind of interference can be reduced if *receiver-directed bit-homogeneous* code assignment is used.

It is possible to almost totally eliminate interference from overlapping packets by using a *bit-by-bit code changing* method, and equipping the receiver with a programmable matched filter which follows the pattern as it varies from bit to bit. If the pattern is long enough so that it does not repeat itself during the transmission of the packet, and the packets arrive with at least a few chip times of relative delay, then no interference will ever take place. This truly approaches *perfect capture*. A similar level of capture can also be achieved by assigning orthogonal codes to nodes which the latter use to encode their packets when transmitting. We refer to this as the *transmitter-directed code assignment*. The preamble must contain information regarding the spreading waveform used, thus allowing the receiver to program its matched filter accordingly.

Given the form of modulation and code assignment, the *SELECTION OF A CHANNEL ACCESS PROTOCOL* must be made. Similar to that of narrow-band systems, we distinguish two types; the ALOHA type and the activity-sensing type. The ALOHA schemes are identical to those in narrow-band system (transmit as long as not already transmitting, nor locked onto a packet worth receiving). A protocol of the activity-sensing type, on the other hand, may or may not be feasible depending on the ability for a node to dynamically acquire knowledge regarding the state of other nodes. Consider CSMA for example. A node must be able to sense activity due to its neighbors. This is possible with space-and-bit homogeneous codes, and is done by observing the output of the matched filters corresponding to the desired waveforms. In transmitter-directed or receiver-directed bit-homogeneous code assignments, a node will have to possess a bank of filters matched to all possible codes used by the neighbors. While this is clearly possible, it is rather impractical. With bit-by-bit code changing, activity sensing is difficult to achieve. From another point of view, namely that of overall network performance, it is not definitely clear that CSMA actually provides any improvment, since spread spectrum already exhibits strong capture properties, and inhibiting transmissions may actually decrease network throughput.

In particular situations one may be able to devise simple but useful schemes. Consider, for example, the case of receiver-directed code assignment, and let the node wishing to transmit monitor the channel for transmissions using the code assigned to the intended destination. If activity is sensed, then it is likely that the intended destination is busy, locked onto a packet destined to it. The existence of hidden nodes, and the possibility that the intended receiver is free (not locked) despite the presence of activity, introduces complexity similarly to that of a narrow-band system.

In spread-spectrum systems, performance is often not the only issue. Because the use of spread spectrum is often driven by operational requirements, considerations such as security, feasibility of implementation, cost, etc., must also be taken into account.

D. Data Link Control

Data link control pertains to the functions at the data link Layer which achieve reliable communications between adjacent nodes (i.e., nodes which are connected directly by a digital radio link). (The relation of link reliability to end-to-end reliability is discussed below when network management and routing are discussed.) As in any packet communication system where some degree of reliability across links is needed, acknowledgment mechanisms (ARQ) are typically used to notify a device of its success in the transmission of a packet. In packet radio networks, however, where the performance of a digital link is highly variable (due to radio propagation characteristics and user contention), and at times poor, acknowledgment procedures alone may not be sufficient and have to be augmented by forward error correction (FEC) coding. Indeed, if the likelihood of errors in a packet is high, then ARQ schemes would result in very low throughput, as most of the packets would be rejected. FEC would greatly improve the chances of correct reception. This is particularly important when spread spectrum is being used with a pseudo-random generation of the code on a bit-by-bit basis.[1] The probabilities of generating two codes that have high correlation sometime during the packet (on two different simultaneous transmissions) is often significant, and therefore the probability of incurring at least a few errors during the packet can be very high. Forward error correction can be used to correct these few bit errors so that the effective packet probability of error can be driven down to an acceptable level for an ARQ scheme to be effective. The rate of coding for error correction, on the other hand, should not be too low, since the information throughput across the link would then be low.

Thus the primary issue in data link control is HOW TO COMBINE FEC AND ARQ SO AS TO ACHIEVE AN ADEQUATE LEVEL OF LINK PERFORMANCE in the highly variable conditions that are typical of packet radio network environments. We note here that the balance between the two techniques must vary from link to link, and dynamically in time to match the current conditions.

The second issue to be addressed at the data link control level is HOW TO IMPLEMENT HOP-BY-HOP ACKNOWLEDGMENTS. One alternative is to have the receiving node transmit *explicit* short acknowledgments consisting typically of only the header, since the header uniquely identifies the packet. Another alternative, due to the local broadcast property of a radio channel, allows the relaying of a packet by the next node to be the acknowledgment to the current node. This scheme is refered to as the *echo or*

[1] The reader may wish to contemplate the view of spread spectrum as a form of FEC.

passive acknowledgment scheme. Clearly, in this scheme, the acknowledgment at the last hop has to be explicit.

While echo acknowledgments might appear to save channel time, as compared to sending explicit acknowledgments, specific problems are associated with them. Consider, for example, what happens when nodes are implementing a single first-in-first-out (FIFO) transmit queue with only one outstanding packet awaiting acknowledgment from a neighboring node. Let *A* and *B* be two consecutive nodes on the path for some source–destination pair. Node *A* transmits the packet at the head of its queue to node *B*. When received by node *B*, this packet is put at the bottom of its transmit queue. Then, the echo acknowledgment awaited for by node *A* will not be received until node *B* services that packet, which is at the bottom of its queue. This problem is particularly severe if packets are traveling along a string of repeaters.

Another problem associated with echo acknowledgments is that, since they are as long as the original packet, they are more likely to be interfered with than would be the case with shorter explicit acknowledgment packets. This may cause severe degradation in performance, especially in highly congested regions, since the original node would be required to undertake additional transmissions beyond the first successful one, due to simply having missed the echo acknowledgment. Thirdly, echo acknowledgments cannot be used in spread-spectrum systems with receiver-dependent codes, since the next node would not be using the PN waveform corresponding to the original node. Explicit acknowledgments, on the other hand, seem to present benefits which would outweigh the presumably additional channel time that they would require. In addition to being less prone to interference, they can be given priority over regular packets, thus speeding up the freeing of the buffers.

III. Network Management

Once the methods for getting data from one node to the next have been determined, the next set of issues pertain to the techniques for moving data through the network. Given the particular environment in which a network is to operate, the characteristics of the radio links, the capabilities of the digital processing, and the desired functional capabilities of the network, design choices must be made as to the various network management algorithms and techniques to be used. In this section, we discuss some of the choices that are available and why certain choices are more applicable for different applications.

For convenience, the discussion is broken into three broad areas: link determination and control, routing and packet forwarding, and other network management concerns such as monitoring the status of the various network nodes. Again, as is true for most areas of packet radio network design, it must be kept in mind that these areas are by no means independent and the actual design choices must be made as a coherent whole.

A. Link Determination and Control

In Section II, we discussed how data can be moved at a suitable level of reliability from one node to its adjacent node (the next node along the path to the destination.) However, there are many parameters associated with transmission and it is the responsibility of network management techniques to determine the logical link connectivity and to control the radio parameters to assure that this link connectivity is maintained and managed appropriately.

HOW SHOULD TWO PACKET RADIO UNITS DETERMINE THE EXISTENCE OF A LINK BETWEEN THEM AND PASS THAT INFORMATION TO THE NETWORK MANAGEMENT ALGORITHMS? The network management techniques (to be discussed below) rely on the ability to pass packets/data from one node to the next and to have knowledge about that packet-passing capability. Because of the dynamics in a mobile packet radio network, it is desirable for this determination to be done as quickly as possible. However, there is also the desire to minimize the number of logical connectivity changes due to transient conditions, such as noise or temporary connectivity outages (such as that which might occur when a mobile radio goes under a bridge.)

Radio connectivity must be determined by the two ends of the radio link (i.e., the two packet radio units which are connected). The information from each node can be collected at a central location where connectivity is then determined, or it can be determined by the nodes themselves through a cooperative mechanism, such as exchange of the number of transmitted and received packets. In either case, a decision must be made as to the nature of the information that will be used to determine the existence of a link.

One set of methods that can be used to determine the existence of a link is to directly use measurements made on the radio channel. These can include such measurements as signal strength, signal-to-noise ratio, and bit error rate. These measurements, made on a packet-by-packet basis and associated with the individual radio-to-radio link, can then be integrated over several packets to declare whether a particular link is useable or not.

It is relatively straightforward to include such measurements in a radio [17], [18]; however, they have the disadvantage of requiring suitable hardware to perform the measurements and therefore may drive up the cost of the radio units. Furthermore, some systems are based on the use of existing radios [2] and adding special hardware would be difficult and expensive.

To avoid the use of special hardware as well as make measurements that are directly related to the operation of the link in the network, direct observations of the link logical level can be made. This is commonly done by simply counting the percentage of packets that are received correctly over some period of time.

Early versions of the DARPA packet radio network [19] utilized this method of determining packet radio link quality. The disadvantage of this method is that it requires a measurement interval sufficiently long to obtain a reasonable estimate of correctly received packets. Thus recognition of a change in connectivity will be delayed until such a time interval has passed. This can result in limitations on the speed of tracking for mobile units (as they move out of connectivity with one neighboring unit and into the range of another.)

In addition to assessing link connectivity, network management algorithms must be concerned with connectivity control. As described in Section II, link connectivity at the physical level (and therefore the logical level) depends on several parameters including data rate, coding rate, and transmitted signal power. To the extent that these are variable, link connectivity assessment must account for the variations and permit network management algorithms to exploit these choices. For example, even the earliest DARPA packet radio units had the capability to control transmitted power and data rate [3]. Data could be transmitted at either 100 or 400 kbits/s at a constant spread-spectrum spreading bandwidth. The result is that the lower data rate could be used when connectivity was poor (either due to increased propagation attenuation of multipath). Algorithms were developed to dynamically select between the two data rates on a per-transmission basis, and this markedly increased the performance of the network over using either data rate exclusively. On the other hand, because of the complex interactions between transmitted power and congestion in the network, satisfactory algorithms to dynamically control transmitted power were not available, partly because the necessary signal strength monitoring tools were not in the earlier radios.

WHAT IS THE CORRECT BALANCE OF LINK PARAMETERS, FORWARD ERROR CORRECTION, HOP-BY-HOP ERROR DETECTION AND RETRANSMISSION (ARQ), AND END-TO-END ARQ? At the link level, decisions have to be made as to the mechanism to obtain some level of reliability. This is due to the fact that a typical packet radio channel has a nonnegligible packet error and loss rate. Relying on only end-to-end mechanisms can result in an overly large number of retransmissions. Therefore, by paying some cost at the link level, overall network channel utilization and delay is improved.

In the section above on data link control, the tradeoff between the various link parameters was discussed. In addition, there must be an interaction between network level routing algorithms (discussed below) and the control of the link parameters [20]. If link connectivity is lost, the network must determine whether it should try harder on that link (by, for example, increasing power or coding gain) or it should attempt to find a different route, thereby possibly suffering some delay and lost packets while the new route is determined.

WHEN DOES THE SET OF AVAILABLE LINKS CONSTITUTE AN ACCEPTABLE NETWORK? Once the available links are determined, the network management algorithms need to determine whether or not the links are sufficient for the network algorithms to proceed to constitute a network. For example, in order that the network be robust in coping with failing nodes and links, it may be desirable to consider only networks where there are a minimum of two (or more) neighbors for every node. Another example is that of partitioned networks. It is possible that the set of links does not form a connected graph (i.e., the radios are clustered with no connectivity between the clusters). This could be managed as two separate networks or as a single partitioned network. Finally, there is the issue of minimal supported traffic levels. If there is a minimal user requirement stated for the network, it might be desirable to accept only those combinations of links (and associated capacities) that will support that traffic environment. For example, if two clusters of radios have a high degree of traffic between them (by specification) and a low degree of connectivity, one must decide whether or not to permit degraded operation.

The issue here then is to determine the minimal acceptable network. At one extreme, one can take the "what is given is what you have" approach, and require that the network management algorithms be prepared to cope with the available digital link connectivity, regardless of what it is. The other extreme would involve a substantial amount of pre-deployment engineering and require that a certain minimal amount of that connectivity be supported. Most of the current approaches to packet radio networking have favored the former approach, recognizing that locations of radios are usually determined by other factors than radio connectivity (such as the user location and mission).

B. Routing and Packet Forwarding

The basic job of the network management algorithms is to allow data packets to be routed through the network in an efficient and reliable manner. This entails two basic tasks. The first is the establishment of routes through the network, and the second is the forwarding of packets along those routes.

At this point, we should note that many of the network management algorithms discussed here are used for other networks in addition to packet radio. However, the unique environment of packet radio, having to do with the unpredictable and changing topology coupled with the local broadcast capability of the radio channel, gives rise to a set of concerns in designing the network management strategies that is significantly different than in other networks (such as long-haul wire networks or local area networks).

HOW SHOULD ROUTES BE ESTABLISHED THROUGH A PACKET RADIO NETWORK BASED ON THE BASIC LINK CONNECTIVITY? A route is the set of links traversed by a packet as it proceeds from source packet radio unit to destination unit. To effectively utilize the available links, routes need to be determined. The choices to be made in this area fall into two areas. WHAT TYPE OF ROUTING AND ROUTING ALGORITHM SHOULD BE USED? HOW SHOULD THE ROUTING INFORMATION BE DISSEMINATED?

1) Type of Routing: Because the methods for network management depend heavily on the method for routing of packets, it is important that the type of routing be resolved early in the design of the network. The methods for routing packets fall into two basic categories. The first is flooding techniques, and the second is point-to-point methods.

Flooding methods involve transmitting the packet to every node in the network. No attempt is made to store routes. Rather, nodes keep track of individual packets as they pass through and decide whether or not to retransmit (usually based on whether they have seen the packet previously). The utility of flooding techniques in packet radio networks arises from their utilization of the inherently broadcast nature of the radio channel. The main advantage of flooding techniques is that they usually involve little explicit overhead and require little network management. They are also well suited to distributed control as many such methods do not require any central control at all. On the other hand, flooding methods tend to utilize the network inefficiently, as every node in the network will receive every packet at least once.

Thus flooding methods tend to be well suited to applications where there is a high need for reliable delivery in the presence of uncertain connectivity and when the connectivity is changing so rapidly that it is difficult for routing information to be determined and disseminated throughout the network in a consistent manner. Flooding methods therefore have potential application in two areas of packet radio network management. The first is for environments where connectivity is changing extremely rapidly so it is inefficient or impossible (given the delays in the network) to track changes in connectivity. The second application is in the area of network control itself. Because flooding techniques do not require *a priori* knowledge of the network connectivity, they are easily used for disseminating network management and control information which is used to determine that connectivity.

Point-to-point routing methods typically involve the association of a route (a sequence of links) with a source–destination pair. One method of doing point-to-point routing is to explicitly associate information in each node with a source–destination pair (connection). Typically such techniques involve a route establishment phase that occurs when the "connection" is first recognized, and then the information stored at each node is used to perform the actual routing of the packets. Forwarding of packets then simply involves looking up the appropriate forwarding information based on the connection identifier (which is carried in the packet). If topology changes occur, a new route establishment (or re-establishment) phase would occur to assure that the correct information is stored at all the nodes in the intended route.

Connectionless approaches to routing typically involve the use of routing techniques that take place as a background activity and do not require an explicit route establishment at the time the end-to-end connection (source destination pair) first has traffic. The individual nodes in the network have no knowledge of the existence of an end-to-end connection, and operate based on information contained in the packet network header (such as the destination address and type of service) and information about the network topology that results from the background operation of the network. Thus as topology changes occur, the background activity would cause the nodal information to be updated without regard to any end-to-end connection,[2] and the traffic would keep flowing (except for some possible delays while topology information is out of date.)

The choice of the routing method used depends heavily on the nature of the traffic pattern and the dynamics of the network topology. Connection oriented approaches have the advantage of requiring minimal information in the packet itself (basically just a connection identifier and sequence number) and so lead to better utilization on the channel. Since channel utilization in a radio environment is always an important consideration, such approaches can be attractive.

However, in networks where topology changes rapidly, routing strategies that lead to local adaptive behavior are preferable to connection oriented approaches, which often require re-establishment of the end-to-end connection when any change occurs in the network topology. Connectionless approaches coupled with distributed routing techniques can often deal with topology changes in a way that maintain the end-to-end service.

Thus we see that all three routing methods have a place in packet radio networks. In relatively static networks, it is often most efficient to have the nodes determine their connectivity, and then determine relatively fixed routes (which would then be modified if connectivity changed due to mobility, etc.). For more dynamic networks, where connectivity is constantly changing, higher channel efficiency can be achieved by reducing the connection setups and the associated overhead. Finally, in the most dynamic networks, where network delays preclude tracking of connectivity on any but the most local basis, flooding techniques would appear to be a reasonable approach.

2) Spreading Routing Information: HOW SHOULD THE INFORMATION THAT EACH NODE REQUIRES TO ROUTE PACKETS BE DISSEMINATED TO THOSE NODES? For any type of routing method (with the exception of the most simple flooding methods), the local connectivity information must be processed and made available to the nodes so that they may route the packets. Note that this is somewhat independent of the type of routing being used. However, it does depend on the method for determining link connectivity and in particular, where the resulting connectivity information resides.

A popular method for doing routing in networks where functional distribution is not needed (e.g., for survivability) is to use a centralized routing server. (This, in fact, was the method used in the early DARPA packet radio network [3].) This technique has each node send its local connectivity information to a central location. At this location, routes are determined and the information required by each node to process and forward packets (such as the next node along the route) is sent to the individual network nodes on either a request basis or as a background operation which constantly updates tables in the nodes.

Use of a centralized routing server has several advantages over more distributed techniques. Because the server has all the connectivity information available (albeit not necessarily current), it can be quite efficient in the computation of routes. This can be a significant advantage in packet radio situations where both connectivity and congestion are more visible globally and where some nodes are typically collocated with mobile users as opposed to being located in some predetermined location. The centralized techniques can generally be extended to a small number of servers for load-sharing and/or backup, thus overcoming some of the problems of size and robustness inherent in a centralized method.

Perhaps the major disadvantage of a centralized technique is its limited ability to handle the rapid local topology changes that often are typical in a packet radio network. Because the connectivity information has to travel to the centralized server, and then the resulting routing information has to be disseminated to the required nodes, centralized methods are inherently limited in their ability to deal with rapid changes in topology. Distributed techniques can, if so designed, often deal with such changes on a local basis.

One method for distributing the routing process is to provide enough information to each node so that each node can simply compute for itself the best total route and then take action locally that is commensurate with that global optimum. For example, based on the computed best total route, a node may determine which is the best node to forward the packet. At the next node, the route may be recomputed or the entire route (or portion) could be included in the packet. (The latter is considerably less robust in the

[2] Sometimes use of the connection might cause the information to be updated more rapidly as a side effect.

face of changing topology.) This form of distributed routing can be accomplished by having each node transmit its local connectivity information explicitly to every other node. Typically a form of flooding is used to disseminate the information.

This method is quite robust (except for errors in tables or transmissions) and, in fact, is the (new) algorithm used in the Arpanet [21] and is planned for use in the gateways of the DARPA Internet system [22], [23]. However, if the network has a relatively high rate of topology changes, the amount of traffic on the network could be very high, as every substantial topology change can produce a number of packets roughly equal to the number of nodes in the network times the number of nodes directly affected by the change. Thus this method of routing is well-suited to a network like the Arpanet or a packet radio network consisting of fixed locations where topology changes are infrequent.

Another interesting routing structure occurs when packet radio networks are hierarchically organized. If the network is assumed to consist of clusters of packet radios that are interconnected, the topology between clusters is likely to change at a slower rate than that between radios, and therefore hierarchical techniques may be applicable. We see this applied to packet radio in [24] and [6].

An even more distributed method relies on each node only knowing information relative to local routing decisions. One method for accomplishing this is for each node to inform its neighbors (and only its neighbors) about the current state of its routing table (the table associating destinations or connections with the next node to be used). If that table contains an additive routing metric (such as number of hops to the destination), a neighboring node can determine, based on the contents of the tables that it hears, the metric for its own table and the next node it should use to route the packet.

Such a routing technique is inherently well-suited to deal with the rapid topology changes that can occur in a packet radio network. However, explicit mechanisms may be necessary to deal with robustness issues (such as route loops). This, in fact, was the algorithm used originally in the Arpanet [25] and it was discarded because, in the Arpanet environment where local broadcast is not convenient, it was difficult to avoid some of the problems. The packet radio environment, where the radio channel affords easy broadcast, is more suitable for this algorithm. A more detailed discussion of the above two techniques and their tradeoffs is contained in [26].

Analyses of the tradeoffs between the various routing strategies have indicated considerable sensitivity to the particular assumptions about topology and topological changes. In addition, different routing techniques may be preferable at different levels in the network organization. For example, the current DARPA packet radio algorithms use the last technique described above to do routing inside clusters of radios, and uses a multiple routing server concept to make routing decisions for routing between clusters [24].

3) Packet Forwarding Issues: Once routes are established, packets are then forwarded from node to node. Packet forwarding techniques are intimately tied to the selection of the routing establishment mechanisms and type of routing. In particular, the selection of the routing mechanism in large part governs the method for forwarding of packets through the network. However, there are a number of issues that need to be dealt with explicitly.

WHAT INFORMATION SHOULD BE PASSED FROM NODE TO NODE IN THE PROCESS OF FORWARDING A PACKET? In addition to the user data, a considerable amount of information associated with network management and control flows through the network. Much of this information is associated with the user data packets. For example, in a connection oriented network, each node must retain a pairing of the virtual circuit identifier and the next node, and each packet must contain the virtual circuit identifier to permit routing to take place. For connectionless routing methods, an indication must be given in the packet of the intended destination. Most if not all routing methods need some unique packet identifier so that duplicates can be identified and eliminated. To ensure valid data, some form of error control information (such as a checksum) must be included in the packet. Often some indication of special requirements for the particular packet must be forwarded (such as priority and delay).

Although the tradeoffs seem simple at the surface (virtual circuit methods require less overhead on a per packet basis than connectionless techniques), the cost per information packet must be balanced against the overhead cost in the network and the resulting functionality. In mobile packet radio environments where topologies are changing rapidly, it is often more effective and less wasteful of overall bandwidth to carry more information in the information packets themselves and have less out-of-band control information flowing.

HOW SHOULD ALTERNATE STRATEGIES BE USED LOCALLY TO TRY TO RAPIDLY CORRECT FOR LOCALIZED TOPOLOGY CHANGES? This issue demonstrates the tight coupling between the various algorithms operating in the dynamic packet radio environment. Most routing and forwarding strategies are designed to work primarily with a relatively fixed route. For example, in a virtual circuit style network, each node has information telling it what next node should receive each packet on a particular virtual circuit. In a connectionless network, a node might have the information to tell it which should be the next node for a particular destination node. However, suppose the connectivity fails between the node and the desired next node. Since this is local information, it is likely that the node will discover this well before global routing tables are updated. Furthermore, it is most likely to discover this while in the process of trying to forward a packet. Therefore there is a possibility of trying to make a local correction to the route based only on local information.

As an example, a strategy that could be used is the following [19]. When a node *A* discovers it cannot reach the desired next node *B*, it sends the user data out in a special packet. This packet is marked "Any node which has connectivity to node *B*, please forward these data to *B*." Thus if a localized rerouting can fix the route, the user data packet can still be delivered in the interim period while the global routing is being repaired. Alternately, no special strategy could be used to reroute the packet, and the node which cannot forward the packet successfully simply would notify the source node as to route failure. The latter has the advantage of being less complex (since it is desirable to fix the global route even if local rerouting can get a packet through).

4) Summary of Routing and Forwarding: As we have seen, there are a large number of tradeoffs involved in the design of the routing and packet forwarding algorithms. Underlying these tradeoffs is a major overall issue; the tradeoff between overhead and responsiveness to changes. Most of the difficulties in the design of the routing algorithms in particular arise from trying to deal with changing topologies, both changes in connectivity and node availability. This is inherent in the packet radio network environment. Certain information has to flow in the network in order to track these changes and respond to them. Yet, the goal of the network is to carry user traffic, not control traffic. Thus the challenge is to balance the need to minimize overhead against the need to track changes in the particular environment of interest.

C. Other Network Management Issues

The above issues dealt with network management and control directly related to the routing of packets. There are a number of other issues that must be dealt with in the management of a network, particularly one with the degree of dynamics associated with a packet radio network.

1) Congestion and Flow Control: HOW CAN TRAFFIC BE LIMITED AT ENTRY TO THE NETWORK AND WITHIN THE NETWORK SO THAT NETWORK CONGESTION IS CONTROLLED? Congestion and flow control are difficult issues to deal with in most networks. The dynamics of packet radio combined with the channel sharing provides additional challenges. Because the topology of a packet radio network is constantly changing, it is very difficult to do "traffic engineering" on the network to assure that it is capable of supporting the network traffic load at all times, even when the traffic originating at each node is limited to a predetermined value (unless, of course, that value is set way below the network capability). Even determining the capacity of a network (the maximum total throughput that a network can pass) given the topology and traffic patterns is difficult [12].

Virtual circuit techniques lead to somewhat direct flow control techniques, as resource reservation can take place in the process of setting up the route. Therefore, if allocated resources are never allowed to exceed that available, there is some assurance that network resources will not be overtaxed. Connectionless and distributed approaches have more technical challenges here. One approach is to detect the presence of increasing congestion (by detecting delay in packet forwarding) and delay packet forwarding based on such detected congestion [27]. Thus the delay tends to propagate through the network back towards the traffic sources and reduce the traffic through the network. This technique is particularly useful in packet radio because of the local broadcast property, and therefore the ability of neighboring radios to detect the current state of packet forwarding delay (particularly if the unit's delay parameter is included in the packet header). Considerable research is still needed in this area, though.

2) Management of Supported Users: In addition to management and control of the internal network units, it is also necessary to manage the interface to the attached user devices.

HOW CAN THE ASSOCIATION OF USER DEVICE TO PACKET RADIO UNIT BE MAINTAINED THROUGHOUT THE NETWORK? When a packet arrives from a user device (or gateway to another network), it typically is marked with the desired destination device. Thus an association must be made between that destination device and the actual packet radio unit to which the device is attached. While this is a problem in all networks, the dynamics of packet radio networks plus their typical operating environment makes this problem particularly severe. As nodes fail, users must have the capability of connecting their devices to replacement radios, and that implies a dynamic name to address association.

One way to do this is to use a static association that can be changed through a manual process. That is, at installation time, the user device is associated with the packet radio unit, and that information is made known to all interested user devices (or a centralized table storing such information). If the device has to be moved to a different packet radio unit (because the unit failed, for example), the process is repeated. The obvious disadvantage to such an approach is the delay in propagating the information to all user devices and, therefore, delay before the user devices are able to communicate.

A preferred approach is to form the association dynamically. When a user device is attached (or detached) from a packet radio unit, the unit detects the fact, determines the user device identification, and passes that information on to either a centralized server or other units using a distributed algorithm [4]. Thus within minutes (or seconds) of moving a user device to another packet radio unit, communications is again possible.

IV. The Operation and Management of a Packet Radio Network

In designing a packet radio network, compatible operation within the data transport and electromagnetic environments must be assured. Some of the issues to achieve this involve design options important to the network users. Some are imposed by constraints such as the radio spectrum and still other issues stem from the ease with which the network is to be operated and maintained.

The data transport environment comprises the various interconnections that join the subject packet radio network to other networks with which connectivity is desired. These collective interconnections are often loosely referred to as the internetwork community or simply the internet. The electromagnetic environment consists of other radio and noise emissions that might adversely impact the packet radio network, and likewise the way in which the packet radio network may have impact on other radio systems.

The perspective in this section is usually that of the network implementor or operator. Occasionally, the role of the user will be examined and some viewpoints will be those of the public interest or the concern of other parties who might be impacted by the operation of such a network.

A. Network Deployment and Maintenance

Critical aspects of packet radio network operation are deployment and maintenance. Deployment is the process of defining the initial topology in a way that meets coverage and capacity requirements, an especially important aspect of mobile operation. Maintenance, as with any distributed system, involves reliability, site visits, and the convenience of repair.

1) Network Deployment: The ease of deployment of a packet radio network is more critical in a military context than a domestic one (although national emergencies often generate requirements similar to those of military deployments). In a military system where deployments may have to occur quickly, some of the resources in each node are devoted to siting aids. But some deployment parameters are common to any system; for example, the level of homing (the number of nodes within radio contact of a given node) and the area covered by each node.

Selecting these parameters constitutes a tradeoff between adequate coverage, redundancy, the spectral conflict imposed by the density of nodes, and, of course, cost. So, an early design issue in deployment is WHAT IS THE APPROPRIATE LEVEL OF HOMING?

Another important aspect of deployment is the degree of automation employed. This can range from the use of built-in siting aids to assist in manual deployment to the total automation of deployment as with the DARPA packet radio system [4]. A very convenient siting aid is an output from the radio that, when the radio is being situated, enumerates the nodes that are its neighbors. In this way, the degree of connectivity can be gauged at the time of network deployment. So another deployment issue is THE DEGREE OF AUTOMATION NEEDED IN THE NETWORK DEPLOYMENT PROCESS.

For example, HOW SHOULD NEW SOFTWARE BE DISSEMINATED TO THE PACKET RADIO UNITS? To avoid having to physically contact units to upgrade software, it is desirable to support downline loading of new software over the network. This is particularly important in a network configuration having unattended repeaters (to overcome obstacles such as mountains). There are several methods for doing this. One is to have a "software distribution server" deliver new software to auxiliary memory in each unit via a network connection. (It is assumed that sufficient software would exist in the unit's ROM (read only memory) to permit network software delivery.)

A more distributed approach is to allow units to load new software from a neighboring unit. The latter approach has the advantage of not requiring any overt action when a new node appears (for example, after being out of radio connectivity). It is more difficult in this case, though, to deal with software integrity and related issues.

In either case, assurance has to be obtained that the nodes are running the most current software. This is usually done by using a version number embedded in the software. This version number also can permit nodes to recognize that a neighboring node is running a more recent version of software, and to request a downline load of the new version.

HOW SHOULD THE VARIOUS PARAMETERS IN A PACKET RADIO UNIT BE SET AND CHANGED? Some parameters in the network are of primarily local concern (such as transmitted power and coding rate) and, therefore, can be set in coordination with the other local radio units. Other parameters, such as frequency, have to be set in coordination with the entire network. Again, a centralized approach can be used with a network control center making any changes from the default settings. Distributed approaches are more difficult, but may lead to increased network functionality.

2) Design Issues Relevant to Maintenance: The topic of maintenance will only be addressed in a limited way here and aspects that cover the design of hardware to make it intrinsically reliable will not be mentioned at all. Two novel approaches that rely on other network resources to assist in the maintenance process are briefly mentioned.

One design characteristic that leads to greater ease of maintenance is the use of common hardware and software in each network node. This assumes that the required range of network functions can be embodied in a single piece of nodal hardware and that the concomitant software differs only in its execution and not in the line-by-line comparison of code. Having this capability not only eases the practice of repair-by-replacement, but also opens the way for internodal network assistance—common resident automatic error detection and cross-net downline loading from a neighbor node. The relevant issue, then, is WHAT ARE THE PERFORMANCE AND COST INEFFICIENCIES OF USING COMMON SOFTWARE AND HARDWARE IN EACH NODE?

The notion of downline loading to repair a software fault or an intermittent or temporary hardware error was mentioned above. This "repair" can be effected automatically or by manual intervention. As in the initialization phase, attention has to be paid to the impact that such downline loading has on normal network traffic. However, the use of such techniques adds to the resiliency of the nodal operation and to the ease of network maintenance.

A network that has been given a wide range of functionality, including adapting to the loss or gain of nodes, complements the hardware in its reliability role. Thus the designer must decide HOW TO ACHIEVE THE HIGHEST LEVEL OF OVERALL NETWORK RELIABILITY FOR A GIVEN COST BY TRADING OFF HARDWARE RELIABILITY AND THE USE OF NETWORKING FEATURES SUCH AS DISTRIBUTED FUNCTIONALITY AND REDUNDANCY?

Evident by this time is that the networking in packet radio systems is not limited to simple store-and-forward transport. Networking defines a more complete, collaborative role among the nodes to accomplish a variety of goals. Obviously, the above issue is a case in point. The ability to diagnose faults in a given or neighbor node and either report such problems or take remedial action enhances the reliability of the system as long as the network does not get "captured" by such internal servicing. Clearly, along with the intelligence to recognize such faults and to attempt repair, comes also the intelligence to devote only so much resource to the task, cauterizing the problem after a specified level of effort.

Redundancy-of-coverage (multiple homing) enhances reliability regardless of the reason for nodal failure. Normally such redundancy mitigates against temporary propagation outages but any nodal failure is compensated for. Too much redundancy in a broadcast system, of course, leads to inordinate collisions and network inefficiency [10].

3) Diagnostics and Monitoring: Diagnostics and monitoring of operation are necessary to the successful operation of a packet radio network. Both functions constitute measurements and can be active, as in the case of probes, or passive, as in traffic monitoring. Both have their impact on network performance through the use of processor time in the switches and the use of air time in the transmitting of probe or reporting packets. The issues, therefore, concern the number of measurements needed, how frequently they are made, and the degree of their passivity.

Monitoring can conveniently be divided into characterizing the operation of individual nodes (switches) and depicting the performance of the nodes collectively. In the first case, functions like buffer occupancy, processing delay, and node throughput are important. In the second case, functions such as routing and the components of network delay are obtained. Monitoring of this type is used in routing and congestion control and was addressed earlier.

But that same monitoring is also important to network operation and maintenance. Short-term problems at a node are dealt with by network management methods such as the temporary halting of routes through a congested node. Longer term congestion or reliability problems must naturally fall out of the monitoring process to be able to invoke the corrective actions of repair and restoration.

HOW SHOULD NODES DETECT WHETHER THEY AND THEIR NEIGHBORS ARE OPERATIONAL AND WHO SHOULD BE NOTIFIED IN CASE OF FAILURE? This is a particularly tricky problem in a mobile packet radio network. A unit can certainly run self-diagnostics to determine its own status. The problem comes in determining if a node has failed totally or has simply moved out of range. Similarly to the routing issues above, there are two strategies for dealing with this. The first is to have a centralized node responsible for keeping track of the existence and status of all nodes in the network. The second is to take a distributed approach, where each node keeps track of either all or a subset of the network nodes.

Neither of these approaches solves the problem of detecting the difference between a failed node, for which some repair action might be needed, and a node which has simply moved out of range of all other nodes in the network, for which the action requried is to either relocate the node or put in place additional repeater nodes. Note, however, that in both of these cases the action required must take place by means outside the normal network operation. (If a node has moved out of range, someone must move it back in range. If a node has failed, someone must repair it.) Thus it would appear to not be unreasonable to relegate this to a local and manual operation, having the operator simply run a diagnostic package when his unit is out of contact with the network.

Other than the manual probing done at the recognition of a problem, the use of diagnostics in the packet radio network occur at power-up. Each node can run self-diagnostics at that time and take appropriate action if tests fail. Examples of built-in testing are the scanning of memory, the parity checking of code as it loads, the cycling of transmit frequencies, and the stepping through the available power levels. In spite of the usefulness of internal tests, the value of cross-net diagnostics and debugging cannot be overemphasized. It too should be part of the network design.

B. Connecting the Packet Radio Network to the External World

Packet radio networks can be operated autonomously or can be connected with other packet-switched networks. The two major areas of concern in the latter case are the specifics of the interconnections process, normally embodied in gateways, and the addressing scheme.

1) Gateways: Gateways can perform many functions but, as far as addressing is concerned, they are packet translation devices that interpret addresses at the internet level and impose headers (addresses) appropriate both to the local networks to which they are attached as well as other networks. They are host-level devices and to work correctly must have some relationship with not only the other gateways of the internet but the network-attached hosts themselves.

Gateways may have an additional role in highly mobile networks such as packet radio where topological partitioning may occur dynamically. Under these circumstances, the gateways, normally internet devices, may take on a role of intranetwork addressing and routing. Specifically, the internet may become the trajectory over which an intranet packet gets delivered when a single network temporarily divides [22]. Whenever gateways play important roles such as this in mobile packet radio networks, the following issue arises: SHOULD ADDRESSING AND ROUTING BE NETWORK- OR GATEWAY-BASED? Network-based addressing means that each network has a unique name and address of which all relevant gateways are aware. In this case, all points within a single network share some portion of their address in common. In contrast, if gateway-based addressing is used, then internet packets are routed from gateway to gateway and each gateway attached to a network must have some means to route packets to destinations within that network. Furthermore, in this case, hosts must have a means to bind themselves dynamically to at least one gateway. Gateway-based routing, while somewhat less intuitive, provides a solution to the problem of what to do when a single network becomes partitioned.

2) Network Access—Methods and Administration: Network access means the functional entry of a network by a person or device capable of using resources within the network or its attached devices. As a general rule it is prudent, depending upon the threat, to exercise access control at the periphery of the network rather than at some centralized (or interior) point or points. Exercising access control at some internal point means that the network must offer a petitioner transport to that point without knowledge as to whether he is entitled to entry or not.

Packet radio with mobile nodes means that access can occur virtually any place within the topology. If so, how can access control best work? Most packet networks provide access through either a connected host or directly through a network-based device (such as a dial-up port). The combination is very convenient, principally for the traveling user who might find it difficult to gain access to a host when not in his normal area. Earlier conventions, wherein network access was not critically controlled and control of host access was invoked at the host only, led to considerable vulnerability to both the network and the attached hosts. Because of the wide host accessability once network access had been gained, network-based access points have characteristically been a weak point in protecting networks from unwanted host entry. Network log-on hosts are increasingly the rule where network-based access is afforded and they may be practical depending on how close to the actual point of network entry access gets controlled.

In mobile packet radio network entry can occur at any node. Mobile users may request entry (connection service) at different places at different times or different places at the same time. Obviously, it becomes more difficult to distribute the access authorization in this situation than if entry were at a fixed location. If access control is decentralized, all nodes may require all authorizations for all mobile users at all times.