Network Interface Guide

Sun Microsystems, Inc.
901 San Antonio Road
Palo Alto, CA 94303-4900
U.S.A.

Part Number 806-1017–10
February 2000

Contents

Preface

The SunOS™ 5.8 Network Interfaces Programmer's Guide describes the basic facilities to implement distributed applications, and guides the programmer in the use of these facilities.

All utilities, their options, and library functions in this manual reflect SunOS Release 5.8. SunOS 5.8 is a new operating system release developed by Sun Microsystems Inc. If you are using a different version of SunOS, some utilities and library functions might function differently.

Audience

A programmer who must convert an existing single-computer application to a networked, distributed application, design a distributed application, implement a distributed application, or maintain a distributed application on the SunOS 5.8 operating system platform should read this manual. Additional techniques for networked applications are described in *ONC+ Developer's Guide*. This manual assumes basic competence in programming, a working familiarity with the C programming language, and a working familiarity with the UNIX operating system. Previous experience in network programming is helpful, but is not required to use this manual.

Organization of the Manual

The services and capabilities of the Network Interfaces portion of the SunOS 5.8 platform are described in the following pages.

Chapter 1 describes this manual and its intent.

Chapter 2 describes the socket interface at the transport layer.

Chapter 3 describes the UNIX System V System Transport Level Interface.

Chapter 4 describes the network selection mechanisms used by applications to select a network transport and its configuration.

Appendixes

Appendix A describes UNIX family sockets.

Appendix B contains complete, functional listings of the code included in the document as examples. These modules are furnished as examples under the provision stated at the beginning of the appendix.

Ordering Sun Documents

Fatbrain.com, an Internet professional bookstore, stocks select product documentation from Sun Microsystems, Inc.

For a list of documents and how to order them, visit the Sun Documentation Center on Fatbrain.com at `http://www1.fatbrain.com/documentation/sun`.

Accessing Sun Documentation Online

The docs.sun.com[SM] Web site enables you to access Sun technical documentation online. You can browse the docs.sun.com archive or search for a specific book title or subject. The URL is `http://docs.sun.com`.

What Typographic Conventions Mean

The following table describes the typographic changes used in this book.

TABLE P–1 Typographic Conventions

Typeface or Symbol	Meaning	Example
`AaBbCc123`	The names of commands, files, and directories; on-screen computer output	Edit your `.login` file. Use `ls –a` to list all files. `machine_name% you have mail.`
`AaBbCc123`	What you type, contrasted with on-screen computer output	`machine_name%` **`su`** `Password:`
AaBbCc123	Command-line placeholder: replace with a real name or value	To delete a file, type **`rm`** *filename*.
AaBbCc123	Book titles, new words, or terms, or words to be emphasized.	Read Chapter 6 in *User's Guide.* These are called *class* options. You must be *root* to do this.

Shell Prompts in Command Examples

The following table shows the default system prompt and superuser prompt for the C shell, Bourne shell, and Korn shell.

TABLE P–2 Shell Prompts

Shell	Prompt
C shell prompt	`machine_name%`
C shell superuser prompt	`machine_name#`

TABLE P–2 Shell Prompts *(continued)*

Shell	Prompt
Bourne shell and Korn shell prompt	$
Bourne shell and Korn shell superuser prompt	#

CHAPTER 1

Introduction to Network Interfaces

This manual describes the programmer's interface to network services in the SunOS 5.8 operating system.

SunOS 5.8 is fully compatible with System V, Release 4 (SVR4) and conforms to the third edition of the System V Interface Description (SVID). It supports all System V network services.

Networking in SunOS 5.8

The theme of networking in SunOS 5.8 is transport independence. Networked applications can execute without having to be tailored to a specific transport protocol.

Previous versions of the system contain sockets, TLI, and name-to-address translation functions. In SunOS 5.8 these are enhanced and work with the new network selection facility to free user applications of the details of specific protocols and address formats.

Transport independent RPC provides interfaces that let applications be free of or more closely tied to the underlying transport. It is the developer's choice to use the most appropriate level.

Applications that must adjust options or use specific addresses can still do so. But you can now write applications to be very portable over different protocol stacks.

Another important feature of SunOS 5.8 is standardized internal kernel network interfaces at the transport and link levels. At the transport level, the AT&T Transport Provider Interface is required. At the link level, the UNIX International Data Link Provider Interface is required.

Standardizing on these interfaces lets you interchange STREAMS drivers at the transport and link level with no changes to the modules or drivers communicating with them. In particular, TLI and sockets can interface to any transport provider supporting TPI, and any device driver supporting DLPI can be linked beneath the Internet Protocol (IP).

Open Systems Interconnect Reference Model

The Open Systems Interconnect (OSI) reference model is the basis of commercially available network service architectures. Other network protocols, developed independently, conform loosely to the model. The TCP/IP Internet Protocol suite is an example.

The OSI reference model is a convenient framework for networking concepts. Basically, data are injected into a network by a sender. The data are transmitted along a communication connection and are delivered to a receiver. To do this, a variety of networking hardware and software must work together.

The OSI reference model divides the functions of networking into seven layers, as depicted in Figure 1-1.

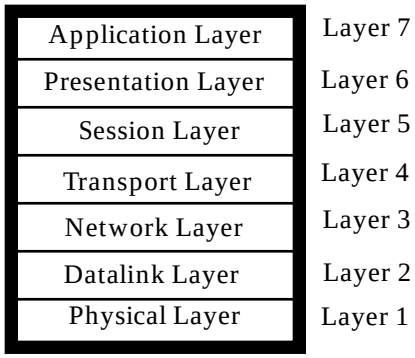

Figure 1–1 OSI Reference Model

Each protocol layer performs services for the layer above it. The ISO definition of the protocol layers provides designers some freedom of implementation. For example, some applications skip the presentation and session layers to interface directly with the Transport layer.

OSI Layer Definitions

Layer 1: Physical Layer

The hardware layer of the model. On SPARC™ systems, it consists of the connector to the network transmission medium, any multiplexor boxes, and cables.

Layer 2: Data Link Layer

Does the sending and receiving. On the sending end, Ethernet (or similar) software organizes the data into packets of appropriate size and packages them. The packaging includes the physical address of the intended receiver. The layer also transmits the message packets and retransmits them if needed.

On the receiving end, the Ethernet hardware recognizes packets with its address and receives them. The Ethernet software strips the transmission packaging and reassembles the data. It can detect transmission errors.

Layer 3: Network Layer

Does the message routing, including translation from logical to physical addresses. The Internet Protocol (IP) is the normal network layer for SPARC systems.

Layer 4: Transport Layer

Controls the flow of data on the network. In SunOS 5.8, any of the Transport Layer Interface (TLI), the Transmission Control Protocol (TCP), or the User Datagram Protocol (UDP) can be used. In SPARC systems, connection mode service is typically provided through TCP, and connectionless service is typically provided through UDP.

Layer 5: Session Layer

Manages reliable sessions between processes. Remote Procedure Calls (RPC) belong at this layer. The interface at this layer allows remote communication using function call semantics.

Layer 6: Presentation Layer

Performs the translation between the data representation local to the computer and the processor-independent format that is sent across the network. In the SunOS 5.8 environment, the processor-independent data format is XDR.

Layer 7: Application Layer

At this top layer are the user-level programs and services. Examples of user-level programs are `telnet`, `rlogin`, `ftp`, and `yppasswd`. Examples of services are NFS™, NIS, and DNS.

Industry standards have been or are being defined for each layer of the reference model. Two standards are defined for each layer: one specifies the interface to the services provided by the layer, and the other specifies the protocol observed by the services in the layer. Users of a service interface standard should be able to ignore the protocol and any other implementation details of the layer.

The Transport Layer

The transport layer (layer 4) is the lowest layer of the model that provides applications and higher layers with end-to-end service. This layer hides the topology and characteristics of the underlying network from users. The transport layer also defines a set of services common to many contemporary protocol suites including the ISO protocols, Transmission Control Protocol and TCP/IP Internet Protocol Suite, Xerox Network Systems (XNS), and Systems Network Architecture (SNA).

In RPC programming, the term “network” is frequently used as a synonym for transport or transport type.

Transport Layer Interface

The Transport Layer Interface (TLI) is modeled on the industry standard Transport Service Definition (ISO 8072). It also can be used to access both TCP and UDP. It is implemented as a user library using the STREAMS I/O mechanism.

CHAPTER 2

Socket Interfaces

This chapter presents the socket interface and illustrates it with sample programs. The programs demonstrate the Internet family sockets.

Sockets Are Multithread Safe

The interface described in this chapter is multithread safe. Applications that contain socket function calls can be used freely in a multithreaded application. Note, however, that the degree of concurrency available to applications is not specified.

SunOS 4 Binary Compatibility

Two major changes from SunOS 4 hold true for SunOS 5 releases. The binary compatibility package allows SunOS 4–based dynamically linked socket applications to run on SunOS 5.

1. You must explicitly specify the socket library (`-lsocket` or `libsocket`) on the compilation line.

2. You may need to link with `libnsl` also (use `-lsocket -lnsl`, not `-lnsl -lsocket`).
3. You must recompile all SunOS 4 socket-based applications with the socket library to run under SunOS 5.

Overview of Sockets

Sockets are the most commonly used low-level interface to network protocols. They have been an integral part of SunOS releases since 1981. A socket is an endpoint of communication to which a name can be bound. A socket has a *type* and one associated process. Sockets were designed to implement the client-server model for interprocess communication where:

- The interface to network protocols needs to accommodate multiple communication protocols, such as TCP/IP, Xerox internet protocols (XNS), and UNIX family.
- The interface to network protocols needs to accommodate server code that waits for connections and client code that initiates connections.
- It also needs to operate differently, depending on whether communication is connection-oriented or connectionless.
- Application programs might want to specify the destination address of the datagrams it delivers instead of binding the address with the `open`(2) call.

Sockets make network protocols available, while behaving like UNIX files. Applications create sockets when they are needed. Sockets work with the `close`(2), `read`(2), `write`(2), `ioctl`(2), and `fcntl`(2) interfaces, and the operating system differentiates between the file descriptors for files and the file descriptors for sockets.

Socket Libraries

The socket interface routines are in a library that must be linked with the application. The library `libsocket.so` is contained in `/usr/lib` with the rest of the system service libraries. `libsocket.so` is used for dynamic linking.

Socket Types

Socket types define the communication properties visible to a user. The Internet family sockets provide access to the TCP/IP transport protocols. The Internet family is identified by the value `AF_INET6`, for sockets that can communicate over both IPv6 and IPv4. The value `AF_INET` is also supported for source compatibility with old applications and for “raw” access to IPv4.

Three types of sockets are supported:

1. Stream sockets allow processes to communicate using TCP. A stream socket provides bidirectional, reliable, sequenced, and unduplicated flow of data with no record boundaries. After the connection has been established, data can be read from and written to these sockets as a byte stream. The socket type is SOCK_STREAM.
2. Datagram sockets allow processes to use UDP to communicate. A datagram socket supports bidirectional flow of messages. A process on a datagram socket can receive messages in a different order from the sending sequence and can receive duplicate messages. Record boundaries in the data are preserved. The socket type is SOCK_DGRAM.
3. Raw sockets provide access to ICMP. These sockets are normally datagram oriented, although their exact characteristics are dependent on the interface provided by the protocol. Raw sockets are not for most applications. They are provided to support developing new communication protocols or for access to more esoteric facilities of an existing protocol. Only superuser processes can use raw sockets. The socket type is SOCK_RAW.

See “Selecting Specific Protocols” on page 49 for further information.

Interface Sets

SunOS 5.8 provides two sets of socket interfaces. The BSD socket interfaces are provided and, since SunOS 5.7 the XNS 5 (Unix98) Socket interfaces are also provided. The XNS 5 interfaces differ slightly from the BSD interfaces.

The XNS 5 Socket interfaces are documented in the man pages: accept(3XNET), bind(3XNET), connect(3XNET), endhostent(3XNET), endnetent(3XNET), endprotoent(3XNET), endservent(3XNET), gethostbyaddr(3XNET), gethostbyname(3XNET), gethostent(3XNET), gethostname(3XNET), getnetbyaddr(3XNET), getnetbyname(3XNET), getnetent(3XNET), getpeername(3XNET), getprotobyname(3XNET), getprotobynumber(3XNET), getprotoent(3XNET), getservbyname(3XNET), getservbyport(3XNET), getservent(3XNET), getsockname(3XNET), getsockopt(3XNET), htonl(3XNET), htons(3XNET), inet_addr(3XNET), inet_lnaof(3XNET), inet_makeaddr(3XNET), inet_netof(3XNET), inet_network(3XNET), inet_ntoa(3XNET), listen(3XNET), ntohl(3XNET), ntohs(3XNET), recv(3XNET), recvfrom(3XNET), recvmsg(3XNET), send(3XNET), sendmsg(3XNET), sendto(3XNET), sethostent(3XNET), setnetent(3XNET), setprotoent(3XNET), setservent(3XNET), setsockopt(3XNET), shutdown(3XNET), socket(3XNET), and socketpair(3XNET).

The traditional SunOS 5 BSD Socket behavior is documented in the corresponding 3N man pages. In addition, a number of new interfaces have been added to section 3N: freeaddrinfo(3SOCKET), freehostent(3SOCKET), getaddrinfo(3SOCKET), getipnodebyaddr(3SOCKET),

getipnodebyname(3SOCKET), getnameinfo(3SOCKET), inet_ntop(3SOCKET), inet_pton(3SOCKET), See the standards(5) man page for information on building applications that use the XNS 5 (Unix98) socket interface.

Socket Basics

This section describes the use of the basic socket interfaces.

Socket Creation

The socket(3SOCKET) call creates a socket in the specified family and of the specified type.

```
s = socket(family, type, protocol);
```

If the protocol is unspecified (a value of 0), the system selects a protocol that supports the requested socket type. The socket handle (a file descriptor) is returned.

The family is specified by one of the constants defined in sys/socket.h. Constants named AF_*suite* specify the address format to use in interpreting names, as shown in Table 2–1.

TABLE 2–1 Protocol Family

AF_APPLETALK	Apple Computer Inc. Appletalk network
AF_INET6	Internet family for IPv6 and IPv4
AF_INET	Internet family for IPv4 only
AF_PUP	Xerox Corporation PUP internet
AF_UNIX	Unix file system

Socket types are defined in sys/socket.h. These types—SOCK_STREAM, SOCK_DGRAM, or SOCK_RAW—are supported by AF_INET6, AF_INET, and AF_UNIX. The following creates a stream socket in the Internet family:

```
s = socket(AF_INET6, SOCK_STREAM, 0);
```

This call results in a stream socket with the TCP protocol providing the underlying communication. Use the default protocol (the *protocol* argument is `0`) in most situations. You can specify a protocol other than the default, as described in "Advanced Topics" on page 44.

Binding Local Names

A socket is created with no name. A remote process has no way to refer to a socket until an address is bound to it. Communicating processes are connected through addresses. In the Internet family, a connection is composed of local and remote addresses, and local and remote ports. There can never be duplicate ordered sets, such as: `protocol`, `local address`, `local port`, `foreign address`, `foreign port`. In most families, connections must be unique.

The `bind`(3SOCKET) call allows a process to specify the local address of the socket. This forms the set `local address`, `local port`. `connect`(3SOCKET), and `accept`(3SOCKET) complete a socket's association by fixing the remote half of the address tuple. The `bind`(3SOCKET) call is used as follows:

`bind (`*s, name, namelen*`);`

The socket handle is *s*. The bound name is a byte string that is interpreted by the supporting protocol(s). Internet family names contain an Internet address and port number.

This example demonstrates binding an Internet address:

```
#include <sys/types.h>
#include <netinet/in.h>
...
struct sockaddr_in6 sin6;
...
  s = socket(AF_INET6, SOCK_STREAM, 0);
  bzero (&sin6, sizeof (sin6));
  sin6.sin6_family = AF_INET6;
  sin6.sin6_addr.s6_addr = in6addr_arg;
  sin6.sin6_port = htons(MYPORT);
  bind(s, (struct sockaddr *) &sin6, sizeof sin6);
```

The content of the address `sin6` is described in "Address Binding" on page 49, where Internet address bindings are discussed.

Connection Establishment

Connection establishment is usually asymmetric, with one process acting as the client and the other as the server. The server binds a socket to a well-known address associated with the service and blocks on its socket for a connect request. An unrelated process can then connect to the server. The client requests services from the

server by initiating a connection to the server's socket. On the client side, the `connect`(3SOCKET) call initiates a connection. In the Internet family, this might appear as:

```
struct sockaddr_in6 server;
 ...
connect(s, (struct sockaddr *)&server, sizeof server);
```

If the client's socket is unbound at the time of the connect call, the system automatically selects and binds a name to the socket. See "Address Binding" on page 49. This is the usual way that local addresses are bound to a socket on the client side.

To receive a client's connection, a server must perform two steps after binding its socket. The first is to indicate how many connection requests can be queued. The second step is to accept a connection:

```
struct sockaddr_in6 from;
 ...
listen(s, 5);                /* Allow queue of 5 connections */
fromlen = sizeof(from);
newsock = accept(s, (struct sockaddr *) &from, &fromlen);
```

The socket handle *s* is the socket bound to the address to which the connection request is sent. The second parameter of `listen`(3SOCKET) specifies the maximum number of outstanding connections that might be queued. `from` is a structure that is filled with the address of the client. A `NULL` pointer might be passed. *fromlen* is the length of the structure. (In the UNIX family, `from` is declared a `struct sockaddr_un`.)

`accept`(3SOCKET) normally blocks. `accept`(3SOCKET) returns a new socket descriptor that is connected to the requesting client. The value of *fromlen* is changed to the actual size of the address.

A server cannot indicate that it accepts connections only from specific addresses. The server can check the from address returned by `accept`(3SOCKET) and close a connection with an unacceptable client. A server can accept connections on more than one socket, or avoid blocking on the accept call. These techniques are presented in "Advanced Topics" on page 44.

Connection Errors

An error is returned if the connection is unsuccessful (however, an address bound by the system remains). Otherwise, the socket is associated with the server and data transfer can begin.

Table 2–2 lists some of the more common errors returned when a connection attempt fails.

TABLE 2–2 Socket Connection Errors

Socket Errors	Error Description
ENOBUFS	Lack of memory available to support the call.
EPROTONOSUPPORT	Request for an unknown protocol.
EPROTOTYPE	Request for an unsupported type of socket.
ETIMEDOUT	No connection established in specified time. This happens when the destination host is down or when problems in the network result in lost transmissions.
ECONNREFUSED	The host refused service. This happens when a server process is not present at the requested address.
ENETDOWN or EHOSTDOWN	These errors are caused by status information delivered by the underlying communication interface.
ENETUNREACH or EHOSTUNREACH	These operational errors can occur either because there is no route to the network or host, or because of status information returned by intermediate gateways or switching nodes. The status returned is not always sufficient to distinguish between a network that is down and a host that is down.

Data Transfer

This section describes the functions to send and receive data. You can send or receive a message with the normal `read`(2) and `write`(2) interfaces:

```
write(s, buf, sizeof buf);
read(s,  buf, sizeof buf);
```

Or the calls `send`(3SOCKET) and `recv`(3SOCKET) can be used:

```
send(s, buf, sizeof buf, flags);
recv(s, buf, sizeof buf, flags);
```

send(3SOCKET) and recv(3SOCKET) are very similar to read(2) and write(2), but the flags argument is important. The flags, defined in sys/socket.h, can be specified as a nonzero value if one or more of the following is required:

MSG_OOB	Send and receive out-of-band data
MSG_PEEK	Look at data without reading
MSG_DONTROUTE	Send data without routing packets

Out-of-band data is specific to stream sockets. When MSG_PEEK is specified with a recv(3SOCKET) call, any data present is returned to the user but treated as still unread. The next read(2) or recv(3SOCKET) call on the socket returns the same data. The option to send data without routing packets applied to the outgoing packets is currently used only by the routing table management process and is unlikely to be interesting to most users.

Closing Sockets

A SOCK_STREAM socket can be discarded by a close(2) function call. If data is queued to a socket that promises reliable delivery after a close(2), the protocol continues to try to transfer the data. If the data is still undelivered after an arbitrary period, it is discarded.

A shutdown(3SOCKET) closes SOCK_STREAM sockets gracefully. Both processes can acknowledge that they are no longer sending. This call has the form:

```
shutdown(s, how);
```

Where how is defined as:

0	Disallows further receives
1	Disallows further sends
2	Disallows both further sends and receives

Connecting Stream Sockets

Figure 2–1 and the next two examples illustrate initiating and accepting an Internet family stream connection.

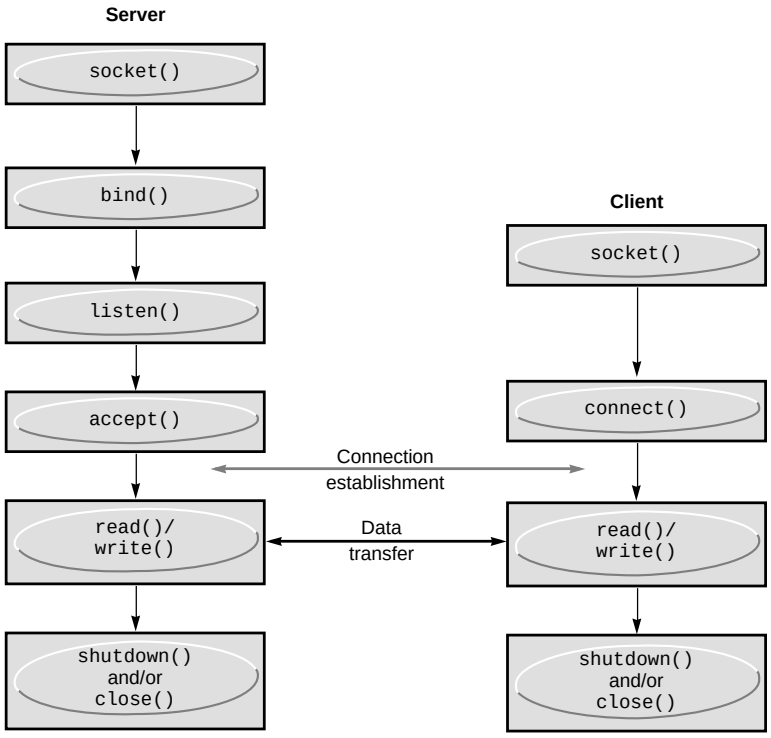

Figure 2–1 Connection-Oriented Communication Using Stream Sockets

The program in Code Example 2–1 is a server. It creates a socket and binds a name to it, then displays the port number. The program calls `listen`(3SOCKET) to mark the socket ready to accept connection requests and initialize a queue for the requests. The rest of the program is an infinite loop. Each pass of the loop accepts a new connection and removes it from the queue, creating a new socket. The server reads and displays the messages from the socket and closes it. The use of `in6addr_any` is explained in "Address Binding" on page 49.

CODE EXAMPLE 2–1 Accepting an Internet Stream Connection (Server)

```
#include <sys/types.h>
#include <sys/socket.h>
#include <netinet/in.h>
#include <netdb.h>
#include <stdio.h>

#define TRUE 1

/*
* This program creates a socket and then begins an infinite loop.
* Each time through the loop it accepts a connection and prints
* data from it. When the connection breaks, or the client closes
* the connection, the program accepts a new connection.
*/
```

```
main()
{
   int sock, length;
   struct sockaddr_in6 server;
   int msgsock;
   char buf[1024];
   int rval;

   /* Create socket. */
   sock = socket(AF_INET6, SOCK_STREAM, 0);
   if (sock == -1) {
      perror("opening stream socket");
      exit(1);
   }
   /* Bind socket using wildcards.*/
   bzero (&server, sizeof(server));
   bzero (&sin6, sizeof (sin6));
   server.sin6_family = AF_INET6;
   server.sin6_addr.s6_addr = in6addr_any;
   server.sin6_port = 0;
   if (bind(sock, (struct sockaddr *) &server, sizeof server)
         == -1) {
      perror("binding stream socket");
      exit(1);
   }
   /* Find out assigned port number and print it out. */
   length = sizeof server;
   if (getsockname(sock,(struct sockaddr *) &server,&length)
         == -1) {
      perror("getting socket name");
      exit(1);
   }
   printf("Socket port #%d\n", ntohs(server.sin6_port));
   /* Start accepting connections. */
   listen(sock, 5);
   do {
      msgsock = accept(sock,(struct sockaddr *) 0,(int *) 0);
      if (msgsock == -1
         perror("accept");
      else do {
         memset(buf, 0, sizeof buf);
         if ((rval = read(msgsock,buf, 1024)) == -1)
            perror("reading stream message");
         if (rval == 0)
            printf("Ending connection\n");
         else
            /* assumes the data is printable */
            printf("-->%s\n", buf);
      } while (rval > 0);
      close(msgsock);
   } while(TRUE);
   /*
    * Since this program has an infinite loop, the socket "sock" is
    * never explicitly closed. However, all sockets will be closed
    * automatically when a process is killed or terminates normally.
    */
   exit(0);
}
```

To initiate a connection, the client program in Code Example 2–2 creates a stream socket and calls `connect`(3SOCKET), specifying the address of the socket for connection. If the target socket exists and the request is accepted, the connection is complete and the program can send data. Data are delivered in sequence with no message boundaries. The connection is destroyed when either socket is closed. For more information about data representation routines, such as `ntohl`(3SOCKET), `ntohs`(3SOCKET), `htons`(3SOCKET), and `htonl`(3XNET), in this program, see the `byteorder`(3SOCKET) man page.

CODE EXAMPLE 2–2 Internet family Stream Connection (Client)

```
#include <sys/types.h>
#include <sys/socket.h>
#include <netinet/in.h>
#include <netdb.h>
#include <stdio.h>

#define DATA "Half a league, half a league . . ."

/*
 * This program creates a socket and initiates a connection with
 * the socket given in the command line. Some data are sent over the
 * connection and then the socket is closed, ending the connection.
 * The form of the command line is: streamwrite hostname portnumber
 * Usage: pgm host port
 */
main(argc, argv)
    int argc;
    char *argv[];
{
    int sock, errnum h_addr_index;
    struct sockaddr_in6 server;
    struct hostent *hp;
    char buf[1024];

    /* Create socket. */
    sock = socket( AF_INET6, SOCK_STREAM, 0);
    if (sock == –1) {
        perror("opening stream socket");
        exit(1);
    }
    /* Connect socket using name specified by command line. */
    bzero (&sin6, sizeof (sin6));
    server.sin6_family = AF_INET6;
    hp = getipnodebyname(AF_INET6, argv[1], AI_DEFAULT, &errnum);
/*
 * getinodebyname returns a structure including the network address
 * of the specified host.
 */
    if (hp == (struct hostent *) 0) {
        fprintf(stderr, "%s: unknown host\n", argv[1]);
        exit(2);
    }

    h_addr_index = 0;
    while (hp->h_addr_list[h_addr_index] != NULL) {
        bcopy(hp->h_addr_list[h_addr_index], &server.sin6_addr,
```

```
                hp->h_length);
        server.sin6_port = htons(atoi(argv[2]));
        if (connect(sock, (struct sockaddr *) &server,
                sizeof (server)) == -1) {
            if (hp->h_addr_list[++h_addr_index] != NULL) {
                /* Try next address */
                continue;
            }
            perror("connecting stream socket");
            freehostent(hp);
            exit(1);
        }
        break;
    }
    freehostent(hp);
    if (write( sock, DATA, sizeof DATA) == –1)
        perror("writing on stream socket");
    close(sock);
    freehostent (hp);
    exit(0);
}
```

Datagram Sockets

A datagram socket provides a symmetric data exchange interface. There is no requirement for connection establishment. Each message carries the destination address. Figure 2–2 shows the flow of communication between server and client.

Note - The `bind`(3SOCKET) step shown below for the server is optional.

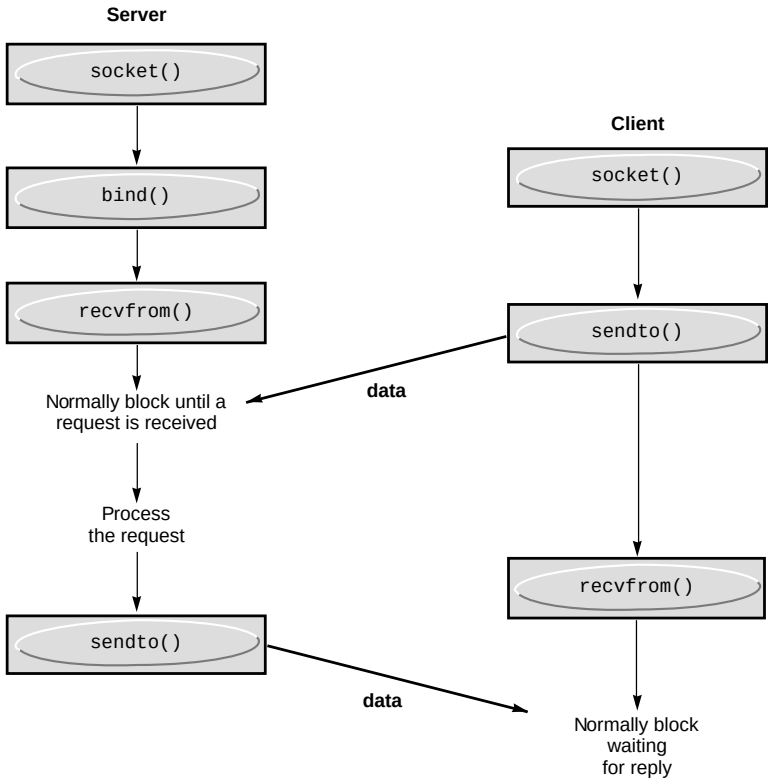

Figure 2–2 Connectionless Communication Using Datagram Sockets

Datagram sockets are created as described in "Socket Creation" on page 20. If a particular local address is needed, the `bind`(3SOCKET) operation must precede the first data transmission. Otherwise, the system sets the local address and/or port when data is first sent. To send data, `sendto`(3SOCKET) is used:

```
sendto(s, buf, buflen, flags, (struct sockaddr *) &to, tolen);
```

The *s*, *buf*, *buflen*, and *flags* parameters are the same as in connection-oriented sockets. The *to* and *tolen* values indicate the address of the intended recipient of the message. A locally detected error condition (such as an unreachable network) causes a return of `–1` and *errno* to be set to the error number.

To receive messages on a datagram socket, `recvfrom`(3SOCKET) is used:

```
recvfrom(s, buf, buflen, flags, (struct sockaddr *) &from, &fromlen);
```

Before the call, *fromlen* is set to the size of the *from* buffer. On return, it is set to the size of the address from which the datagram was received.

Datagram sockets can also use the connect(3SOCKET) call to associate a socket with a specific destination address. It can then use the send(3SOCKET) call. Any data sent on the socket without explicitly specifying a destination address is addressed to the connected peer, and only data received from that peer is delivered. Only one connected address is permitted for one socket at a time. A second connect(3SOCKET) call changes the destination address. Connect requests on datagram sockets return immediately. The system records the peer's address. accept(3SOCKET), and listen(3SOCKET) are not used with datagram sockets.

While a datagram socket is connected, errors from previous send(3SOCKET) calls can be returned asynchronously. These errors can be reported on subsequent operations on the socket, or an option of getsockopt(3SOCKET), SO_ERROR, can be used to interrogate the error status.

Code Example 2–3 shows how to send an Internet call by creating a socket, binding a name to the socket, and sending the message to the socket.

CODE EXAMPLE 2–3 Sending an Internet Family Datagram

```
#include <sys/types.h>
#include <sys/socket.h>
#include <netinet/in.h>
#include <netdb.h>
#include <stdio.h>

#define DATA "The sea is calm, the tide is full . . ."

/*
 * Here I send a datagram to a receiver whose name I get from
 * the command line arguments. The form of the command line is:
 * dgramsend hostname portnumber
 */
main(argc, argv)
   int argc, errnum;
   char *argv[];
{
   int sock;
   struct sockaddr_in6 name;
   struct hostent *hp;

   /* Create socket on which to send. */
   sock = socket(AF_INET6,SOCK_DGRAM, 0);
   if (sock == –1) {
      perror("opening datagram socket");
      exit(1);
   }
   /*
    * Construct name, with no wildcards, of the socket to ‘‘send’’
    * to. getinodebyname returns a structure including the network
    * address of the specified host. The port number is taken from
    * the command line.
    */
   hp = getipnodebyname(AF_INET6, argv[1], AI_DEFAULT, &errnum);
   if (hp == (struct hostent *) 0) {
      fprintf(stderr, "%s: unknown host\n", argv[1]);
      exit(2);
```

```
   }
   bzero (&sin6, sizeof (sin6));
   bzero (&name, sizeof (name));
   memcpy((char *) &name.sin6_addr, (char *) hp->h_addr,
      hp->h_length);
   name.sin6_family = AF_INET6;
   name.sin6_port = htons(atoi(argv[2]));
   /* Send message. */
   if (sendto(sock,DATA, sizeof DATA ,0,
      (struct sockaddr *) &name,sizeof name) == –1)
      perror("sending datagram message");
   close(sock);
   exit(0);
}
```

Code Example 2–4 shows how to read an Internet call by creating a socket, binding a name to the socket, and then reading from the socket.

CODE EXAMPLE 2–4 Reading Internet Family Datagrams

```
#include <sys/types.h>
#include <sys/socket.h>
#include <netinet/in.h>
#include <stdio.h>

/*
 * This program creates a datagram socket, binds a name to it, then
 * reads from the socket.
 */

main()
{
   int sock, length;
   struct sockaddr_in6 name;
   char buf[1024];

   /* Create socket from which to read. */
   sock = socket(AF_INET6, SOCK_DGRAM, 0);
   if (sock == –1) {
      perror("opening datagram socket");
      exit(1);
   }
   /* Create name with wildcards. */
   bzero (&sin6, sizeof (sin6));
   name.sin6_family = AF_INET6;
   name.sin6_addr.s6_addr = in6addr_any;
   name.sin6_port = 0;
   if (bind(sock,(struct sockaddr *)&name, sizeof name) == –1) {
      perror("binding datagram socket");
      exit(1);
   }
   /* Find assigned port value and print it out. */
   length = sizeof(name);
   if (getsockname(sock,(struct sockaddr *) &name, &length)
         == –1)  {
      perror("getting socket name");
      exit(1);
   }
```

```
        printf("Socket port #%d\n", ntohs(name.sin6_port));
        /* Read from the socket. */
        if (read(sock, buf, 1024) == -1 )
           perror("receiving datagram packet");
        /* Assumes the data is printable */
        printf("-->%s\n", buf);
        close(sock);
        exit(0);
    }
```

Input/Output Multiplexing

Requests can be multiplexed among multiple sockets or files. Use `select`(3C) to do this:

```
#include <sys/time.h>
#include <sys/types.h>
#include <sys/select.h>
 ...
fd_set readmask, writemask, exceptmask;
struct timeval timeout;
 ...
select(nfds, &readmask, &writemask, &exceptmask, &timeout);
```

The first argument of `select`(3C) is the number of file descriptors in the lists pointed to by the next three arguments.

The second, third, and fourth arguments of `select`(3C) point to three sets of file descriptors: a set of descriptors to read on, a set to write on, and a set on which exception conditions are accepted. Out-of-band data is the only exceptional condition. Any of these pointers can be a properly cast null. Each set is a structure containing an array of long integer bit masks. The size of the array is set by `FD_SETSIZE` (defined in `select.h`). The array is long enough to hold one bit for each `FD_SETSIZE` file descriptor.

The macros `FD_SET`(*fd*, &*mask*) and `FD_CLR`(*fd*, &*mask*) add and delete, respectively, the file descriptor *fd* in the set `mask`. The set should be zeroed before use, and the macro `FD_ZERO`(&*mask*) clears the set `mask`.

The fifth argument of `select`(3C) allows a time-out value to be specified. If the `timeout` pointer is `NULL`, `select`(3C) blocks until a descriptor is selectable, or until a signal is received. If the fields in `timeout` are set to `0`, `select`(3C) polls and returns immediately.

`select`(3C) normally returns the number of file descriptors selected. `select`(3C) returns a `0` if the time-out has expired. `select`(3C) returns `–1` for an error or interrupt with the error number in *errno* and the file descriptor masks unchanged. For a successful return, the three sets indicate which file descriptors are ready to be read from, written to, or have exceptional conditions pending.

You should test the status of a file descriptor in a select mask with the `FD_ISSET`(*fd*, &*mask*) macro. It returns a nonzero value if *fd* is in the set `mask`, and `0` if it is not.

Use `select`(3C) followed by a `FD_ISSET`(*fd*, &*mask*) macro on the read set to check for queued connect requests on a socket.

Code Example 2–5 shows how to select on a “listening” socket for readability to determine when a new connection can be picked up with a call to `accept`(3SOCKET). The program accepts connection requests, reads data, and disconnects on a single socket.

CODE EXAMPLE 2–5 Using `select`(3C) to Check for Pending Connections

```
#include <sys/types.h>
#include <sys/socket.h>
#include <sys/time.h>
#include <netinet/in.h>
#include <netdb.h>
#include <stdio.h>

#define TRUE 1

/*
 * This program uses select to check that someone is
 * trying to connect before calling accept.
 */

main()
{
   int sock, length;
   struct sockaddr_in6 server;
   int msgsock;
   char buf[1024];
   int rval;
   fd_set ready;
   struct timeval to;

   /* Open a socket and bind it as in previous examples. */

   /* Start accepting connections. */
   listen(sock, 5);
   do {
      FD_ZERO(&ready);
      FD_SET(sock, &ready);
      to.tv_sec = 5;
      to.tv_usec = 0;
      if (select(sock + 1, &ready, (fd_set *)0, (fd_set *)0, &to) == -1) {
         perror("select");
         continue;
      }
      if (FD_ISSET(sock, &ready)) {
         msgsock = accept(sock, (struct sockaddr *)0,
            (int *)0);
         if (msgsock == -1)
            perror("accept");
         else do {
            memset(buf, 0, sizeof buf);
            if ((rval = read(msgsock, buf, 1024)) == -1)
               perror("reading stream message");
            else if (rval == 0)
               printf("Ending connection\n");
```

```
            else
               printf("-->%s\n", buf);
         } while (rval > 0);
         close(msgsock);
      } else
         printf("Do something else\n");
   } while (TRUE);
    exit(0);
}
```

In previous versions of the `select`(3C) routine, its arguments were pointers to integers instead of pointers to `fd_sets`. This style of call still works if the number of file descriptors is smaller than the number of bits in an integer.

`select`(3C) provides a synchronous multiplexing scheme. The `SIGIO` and `SIGURG` signals (described in “Advanced Topics” on page 44) provide asynchronous notification of output completion, input availability, and exceptional conditions.

Standard Routines

You might need to locate and construct network addresses. This section describes the routines that manipulate network addresses. Unless otherwise stated, functions presented in this section apply only to the Internet family.

Locating a service on a remote host requires many levels of mapping before client and server communicate. A service has a name for human use. The service and host names must be translated to network addresses. Finally, the address is used to locate and route to the host. The specifics of the mappings can vary between network architectures. Preferably, a network does not require that hosts be named, thus protecting the identity of their physical locations. It is more flexible to discover the location of the host when it is addressed.

Standard routines map host names to network addresses, network names to network numbers, protocol names to protocol numbers, and service names to port numbers, and the appropriate protocol to use in communicating with the server process. The file `netdb.h` must be included when using any of these routines.

Host and Service Names

The interfaces `getaddrinfo`(3SOCKET), `getnameinfo`(3SOCKET), and `freeaddrinfo`(3SOCKET) provide a simplified method of translating between the names and addresses of a service on a host. For IPv6, these interfaces can be used instead of calling `getipnodebyname`(3SOCKET) and `getservbyname`(3SOCKET) and then figuring out how to combine the addresses. Similarly, for IPv4, these interfaces can be used instead of `gethostbyname`(3NSL) and

getservbyname(3SOCKET). Both IPv6 and IPv4 addresses are handled transparently.

getaddrinfo(3SOCKET) returns the combined address and port number of the specified host and service names. Since all of the information returned by getaddrinfo(3SOCKET) is dynamically allocated, it must be freed by freeaddrinfo(3SOCKET) to prevent memory leaks. getnameinfo(3SOCKET) returns the host and services names associated with a specified address and port number. To print error messages based on the EAI_xxx codes returned by getaddrinfo(3SOCKET) and getnameinfo(3SOCKET), call gai_strerror(3SOCKET).

An example of using getaddrinfo(3SOCKET) follows:

```
struct addrinfo         *res, *aip;
struct addrinfo         hints;
int                     sock = -1;
int                     error;

/* Get host address.  Any type of address will do. */
bzero(&hints, sizeof (hints));
hints.ai_flags = AI_ALL|AI_ADDRCONFIG;
hints.ai_socktype = SOCK_STREAM;

error = getaddrinfo(hostname, servicename, &hints, &res);
if (error != 0) {
    (void) fprintf(stderr, "getaddrinfo: %s for host %s service %s\n",
             gai_strerror(error), hostname, servicename);
    return (-1);
}
```

After processing the information returned by getaddrinfo(3SOCKET) in the structure pointed to by res, the storage should be released by

```
freeaddrinfo(res);
```

getnameinfo(3SOCKET) is particularly useful in identifying the cause of an error as in the following example:

```
struct sockaddr_storage faddr;
int                     sock, new_sock;
socklen_t               faddrlen;
int                     error;
char                    hname[NI_MAXHOST];
char                    sname[NI_MAXSERV];

 ...
    faddrlen = sizeof (faddr);
    new_sock = accept(sock, (struct sockaddr *)&faddr, &faddrlen);
    if (new_sock == -1) {
        if (errno != EINTR && errno != ECONNABORTED) {
            perror("accept");
        }
```

```
            continue;
        }
        error = getnameinfo((struct sockaddr *)&faddr, faddrlen, hname,
                    sizeof (hname), sname, sizeof (sname), 0);
        if (error) {
            (void) fprintf(stderr, "getnameinfo: %s\n",
                        gai_strerror(error));
        } else {
            (void) printf("Connection from %s/%s\n", hname, sname);
        }
```

hostent – Host Names

An Internet host-name-to-address mapping is represented by the `hostent` structure:

```
struct hostent {
    char  *h_name;            /* official name of host */
    char  **h_aliases;        /* alias list */
    int   h_addrtype;         /* hostaddrtype(e.g.,AF_INET6) */
    int   h_length;           /* length of address */
    char  **h_addr_list;      /* list of addrs, null terminated */
};
/*1st addr, net byte order*/
#define h_addr h_addr_list[0]
```

`getipnodebyname`(3SOCKET) maps an Internet host name to a `hostent` structure, `getipnodebyaddr`(3SOCKET) maps an Internet host address to a `hostent` structure, `freehostent`(3SOCKET) frees the memory of a `hostent` structure, and `inet_ntop`(3SOCKET) maps an Internet host address to a displayable string.

The routines return a `hostent` structure containing the name of the host, its aliases, the address type (address family), and a `NULL`-terminated list of variable length addresses. The list of addresses is required because a host can have many addresses. The `h_addr` definition is for backward compatibility, and is the first address in the list of addresses in the `hostent` structure.

netent – Network Names

The routines to map network names to numbers and back return a `netent` structure:

```
/*
 * Assumes that a network number fits in 32 bits.
 */
struct netent {
   char     *n_name;      /* official name of net */
   char     **n_aliases;  /* alias list */
   int      n_addrtype;   /* net address type */
   int      n_net;        /* net number, host byte order */
};
```

getnetbyname(3SOCKET), getnetbyaddr_r(3SOCKET), and getnetent(3SOCKET) are the network counterparts to the host routines described above.

protoent – Protocol Names

The protoent structure defines the protocol-name mapping used with getprotobyname(3SOCKET), getprotobynumber(3SOCKET), and getprotoent(3SOCKET):

```
struct protoent {
   char     *p_name;          /* official protocol name */
   char     **p_aliases       /* alias list */
   int      p_proto;          /* protocol number */
};
```

servent – Service Names

An Internet family service resides at a specific, well-known port and uses a particular protocol. A service-name-to-port-number mapping is described by the servent structure:

```
struct servent {
   char     *s_name;         /* official service name */
   char     **s_aliases;     /* alias list */
   int      s_port;          /* port number, network byte order */
   char     *s_proto;        /* protocol to use */
};
```

getservbyname(3SOCKET) maps service names and, optionally, a qualifying protocol to a servent structure. The call:

```
sp = getservbyname("telnet", (char *) 0);
```

returns the service specification of a telnet server using any protocol. The call:

```
sp = getservbyname("telnet", "tcp");
```

returns the telnet server that uses the TCP protocol. getservbyport(3SOCKET) and getservent(3SOCKET) are also provided. getservbyport(3SOCKET) has an interface similar to that of getservbyname(3SOCKET); an optional protocol name can be specified to qualify lookups.

Other Routines

In addition to address-related database routines, there are several other routines that simplify manipulating names and addresses. Table 2–3 summarizes the routines for manipulating variable-length byte strings and byte-swapping network addresses and values.

TABLE 2–3 Runtime Library Routines

Call	Synopsis
`memcmp`(3C)	Compares byte-strings; `0` if same, not `0` otherwise
`memcpy`(3C)	Copies *n* bytes from *s2* to *s1*
`memset`(3C)	Sets *n* bytes to `value` starting at `base`
`htonl`(3SOCKET)	32-bit quantity from host into network byte order
`htons`(3SOCKET)	16-bit quantity from host into network byte order
`ntohl`(3SOCKET)	32-bit quantity from network into host byte order
`ntohs`(3SOCKET)	16-bit quantity from network into host byte order

The byte-swapping routines are provided because the operating system expects addresses to be supplied in network order. On some architectures, the host byte ordering is different from network byte order, so programs must sometimes byte-swap values. Routines that return network addresses do so in network order. Byte-swapping problems occur only when interpreting network addresses. For example, the following code formats a TCP or UDP port:

```
printf("port number %d\n", ntohs(sp->s_port));
```

On certain machines, where these routines are not needed, they are defined as null macros.

Client-Server Programs

The most common form of distributed application is the client/server model. In this scheme, client processes request services from a server process.

An alternate scheme is a service server that can eliminate dormant server processes. An example is `inetd`(1M), the Internet service daemon. `inetd`(1M) listens at a variety of ports, determined at start up by reading a configuration file. When a connection is requested on an `inetd`(1M) serviced port, `inetd`(1M) spawns the appropriate server to serve the client. Clients are unaware that an intermediary has played any part in the connection. `inetd`(1M) is described in more detail in “ `inetd`(1M) Daemon” on page 59.

Servers

Most servers are accessed at well-known Internet port numbers or UNIX family names. Code Example 2–6 illustrates the main loop of a remote-login server.

CODE EXAMPLE 2–6 Remote Login Server

```
main(argc, argv)
   int argc;
   char *argv[];
{
   int f;
   struct sockaddr_in6 from;
   struct sockaddr_in6 sin;
   struct servent *sp;

   sp = getservbyname("login", "tcp");

   if (sp == (struct servent *) NULL) {
      fprintf(stderr, "rlogind: tcp/login: unknown service");
      exit(1);
   }
   ...
   #ifndef DEBUG
   /* Disassociate server from controlling terminal. */
   ...
   #endif
    sin.sin6_port = sp->s_port; /* Restricted port */
      sin.sin6_addr.s6_addr = in6addr_any;
      ...
      f = socket(AF_INET6, SOCK_STREAM, 0);
      ...
```

```
      if (bind( f, (struct sockaddr *) &sin, sizeof sin ) == –1) {
      ...
      }
      ...
      listen(f, 5);
      while (TRUE) {
         int g, len = sizeof from;
         g = accept(f, (struct sockaddr *) &from, &len);
         if (g == –1) {
            if (errno != EINTR)
               syslog(LOG_ERR, "rlogind: accept: %m");
            continue;
         }
         if (fork() == 0) {
            close(f);
            doit(g, &from);
         }
         close(g);
      }
      exit(0);
}
```

Code Example 2–7 shows how the server gets its service definition.

CODE EXAMPLE 2–7 Remote Login Server: Step 1

```
sp = getservbyname("login", "tcp");
if (sp == (struct servent *) NULL) {
  fprintf(stderr, "rlogind: tcp/login: unknown service\n");
  exit(1);
}
```

The result from `getservbyname`(3SOCKET) is used later to define the Internet port at which the program listens for service requests. Some standard port numbers are in `/usr/include/netinet/in.h`.

Code Example 2–8 shows how the server dissociates from the controlling terminal of its invoker in the non-`DEBUG` mode of operation.

CODE EXAMPLE 2–8 Dissociating From the Controlling Terminal

```
   (void) close(0);
   (void) close(1);
   (void) close(2);
   (void) open("/", O_RDONLY);
   (void) dup2(0, 1);
   (void) dup2(0, 2);
   setsid();
```

This prevents the server from receiving signals from the process group of the controlling terminal. After a server has dissociated itself, it cannot send reports of errors to a terminal and must log errors with `syslog`(3C).

A server next creates a socket and listens for service requests. `bind`(3SOCKET) ensures that the server listens at the expected location. (The remote login server listens at a restricted port number, so it runs as superuser.)

Code Example 2–9 illustrates the main body of the loop.

CODE EXAMPLE 2–9 Remote Login Server: Main Body

```
while(TRUE) {
  int g, len = sizeof(from);
  if (g = accept(f, (struct sockaddr *) &from, &len) == –1) {
   if (errno != EINTR)
    syslog(LOG_ERR, "rlogind: accept: %m");
   continue;
  }
  if (fork() == 0) {  /* Child */
   close(f);
   doit(g, &from);
  }
  close(g);     /* Parent */
}
```

`accept`(3SOCKET) blocks messages until a client requests service. `accept`(3SOCKET) returns a failure indication if it is interrupted by a signal, such as `SIGCHLD`. The return value from `accept`(3SOCKET) is checked and an error is logged with `syslog`(3C) if an error has occurred.

The server then `fork`(2)s a child process and invokes the main body of the remote login protocol processing. The socket used by the parent to queue connection requests is closed in the child. The socket created by `accept`(3SOCKET) is closed in the parent. The address of the client is passed to the server application's `doit()` routine, which performs the actual application protocol with the client, for authenticating clients.

Clients

This section describes the steps taken by the client remote login process. As in the server, the first step is to locate the service definition for a remote login:

```
sp = getservbyname("login", "tcp");
if (sp == (struct servent *) NULL) {
  fprintf(stderr,"rlogin: tcp/login: unknown service");
  exit(1);
}
```

Next, the destination host is looked up by a call to `getipnodebyname`(3SOCKET):

```
hp = getipnodebyname (AF_INET6, argv[1], AI_DEFAULT, &errnum);
if (hp == (struct hostent *) NULL) {
  fprintf(stderr, "rlogin: %s: unknown host", argv[1]);
  exit(2);
}
```

The next step is to connect to the server at the requested host and start the remote login protocol. The address buffer is cleared and filled with the Internet address of the foreign host and the port number at which the login server listens:

```
memset((char *) &server, 0, sizeof server);
bzero (&sin6, sizeof (sin6));
memcpy((char*) &server.sin6_addr,hp->h_addr,hp->h_length);
server.sin6_family = hp->h_addrtype;
server.sin6_port = sp->s_port;
```

A socket is created, and a connection initiated. `connect`(3SOCKET) implicitly does a `bind`(3SOCKET), since `s` is unbound.

```
s = socket(hp->h_addrtype, SOCK_STREAM, 0);
if (s < 0) {
  perror("rlogin: socket");
  exit(3);
}
 ...
if (connect(s, (struct sockaddr *) &server, sizeof server) < 0) {
  perror("rlogin: connect");
  exit(4);
}
```

Connectionless Servers

Some services use datagram sockets. The `rwho`(1) service provides status information on hosts connected to a local area network. (Avoid running `in.rwhod`(1M) because it causes heavy network traffic.) This service requires the ability to broadcast information to all hosts connected to a particular network. It is an example of datagram socket use.

A user on a host running the `rwho`(1) server can get the current status of another host with `ruptime`(1). Typical output is illustrated in Code Example 2–10.

CODE EXAMPLE 2–10 Output of `ruptime`(1) Program

```
itchy up 9:45, 5 users, load 1.15, 1.39, 1.31
scratchy up 2+12:04, 8 users, load 4.67, 5.13, 4.59
click up 10:10, 0 users, load 0.27, 0.15, 0.14
clack up 2+06:28, 9 users, load 1.04, 1.20, 1.65
ezekiel up 25+09:48, 0 users, load 1.49, 1.43, 1.41
dandy 5+00:05, 0 users, load 1.51, 1.54, 1.56
peninsula down 0:24
wood down 17:04
carpediem down 16:09
chances up 2+15:57, 3 users, load 1.52, 1.81, 1.86
```

Status information is periodically broadcast by the `rwho`(1) server processes on each host. The server process also receives the status information and updates a database. This database is interpreted for the status of each host. Servers operate autonomously, coupled only by the local network and its broadcast capabilities.

Use of broadcast is fairly inefficient because a lot of net traffic is generated. Unless the service is used widely and frequently, the expense of periodic broadcasts outweighs the simplicity.

Code Example 2–11 shows a simplified version of the rwho(1) server. It performs two tasks: receives status information broadcast by other hosts on the network and supplies the status of its host. The first task is done in the main loop of the program: Packets received at the rwho(1) port are checked to be sure they were sent by another rwho(1) server process, and are stamped with the arrival time. They then update a file with the status of the host. When a host has not been heard from for an extended time, the database routines assume the host is down and logs it. This application is prone to error, as a server might be down while a host is up.

CODE EXAMPLE 2–11 rwho(1) Server

```
main()
{
   ...
   sp = getservbyname("who", "udp");
   net = getnetbyname("localnet");
   sin.sin6_addr = inet_makeaddr(net->n_net, in6addr_any);
   sin.sin6_port = sp->s_port;
   ...
   s = socket(AF_INET6, SOCK_DGRAM, 0);
   ...
   on = 1;
   if (setsockopt(s, SOL_SOCKET, SO_BROADCAST, &on, sizeof on)
         == –1) {
      syslog(LOG_ERR, "setsockopt SO_BROADCAST: %m");
      exit(1);
   }
   bind(s, (struct sockaddr *) &sin, sizeof sin);
   ...
   signal(SIGALRM, onalrm);
   onalrm();
   while(1) {
      struct whod wd;
       int cc, whod, len = sizeof from;
      cc = recvfrom(s, (char *) &wd, sizeof(struct whod), 0,
         (struct sockaddr *) &from, &len);
      if (cc <= 0) {
      if (cc == –1 && errno != EINTR)
         syslog(LOG_ERR, "rwhod: recv: %m");
      continue;
      }
      if (from.sin6_port != sp->s_port) {
         syslog(LOG_ERR, "rwhod: %d: bad from port",
            ntohs(from.sin6_port));
         continue;
      }
      ...
      if (!verify( wd.wd_hostname)) {
         syslog(LOG_ERR, "rwhod: bad host name from %x",
            ntohl(from.sin6_addr.s6_addr));
         continue;
      }
      (void) sprintf(path, "%s/whod.%s", RWHODIR, wd.wd_hostname);
```

```
            whod = open(path, O_WRONLY|O_CREAT|O_TRUNC, 0666);
            ...
            (void) time(&wd.wd_recvtime);
            (void) write(whod, (char *) &wd, cc);
            (void) close(whod);
        }
        exit(0);
}
```

The second server task is to supply the status of its host. This requires periodically acquiring system status information, packaging it in a message, and broadcasting it on the local network for other `rwho`(1) servers to hear. This task is run by a timer and triggered with a signal. Locating the system status information is involved but uninteresting.

Status information is broadcast on the local network. For networks that do not support broadcast, use another scheme.

Advanced Topics

For most programmers, the mechanisms already described are enough to build distributed applications. Others need some of the additional features in this section.

Out-of-Band Data

The stream socket abstraction includes out-of-band data. Out-of-band data is a logically independent transmission channel between a pair of connected stream sockets. Out-of-band data is delivered independent of normal data. The out-of-band data facilities must support the reliable delivery of at least one out-of-band message at a time. This message can contain at least one byte of data, and at least one message can be pending delivery at any time.

For communications protocols that support only in-band signaling (that is, urgent data is delivered in sequence with normal data), the message is extracted from the normal data stream and stored separately. This lets users choose between receiving the urgent data in order and receiving it out of sequence, without having to buffer the intervening data.

You can peek (with `MSG_PEEK`) at out-of-band data. If the socket has a process group, a `SIGURG` signal is generated when the protocol is notified of its existence. A process can set the process group or process ID to be informed by `SIGURG` with the appropriate `fcntl`(2) call, as described in “Interrupt-Driven Socket I/O” on page 47 for `SIGIO`. If multiple sockets have out-of-band data waiting delivery, a `select`(3C) call for exceptional conditions can be used to determine the sockets with such data pending.

A logical mark is placed in the data stream at the point at which the out-of-band data was sent. The remote login and remote shell applications use this facility to propagate signals between client and server processes. When a signal is received, all data up to the mark in the data stream is discarded.

To send an out-of-band message, the MSG_OOB flag is applied to send(3SOCKET) or sendto(3SOCKET). To receive out-of-band data, specify MSG_OOB to recvfrom(3SOCKET) or recv(3SOCKET) (unless out-of-band data is taken in line, in which case the MSG_OOB flag is not needed). The SIOCATMARK ioctl(2) tells whether the read pointer currently points at the mark in the data stream:

```
int yes;
ioctl(s, SIOCATMARK, &yes);
```

If *yes* is 1 on return, the next read returns data after the mark. Otherwise, assuming out-of-band data has arrived, the next read provides data sent by the client before sending the out-of-band signal. The routine in the remote login process that flushes output on receipt of an interrupt or quit signal is shown in Code Example 2–12. This code reads the normal data up to the mark (to discard it), then reads the out-of-band byte.

A process can also read or peek at the out-of-band data without first reading up to the mark. This is more difficult when the underlying protocol delivers the urgent data in-band with the normal data, and only sends notification of its presence ahead of time (for example, TCP, the protocol used to provide socket streams in the Internet family). With such protocols, the out-of-band byte might not yet have arrived when a recv(3SOCKET) is done with the MSG_OOB flag. In that case, the call returns the error of EWOULDBLOCK. Also, there might be enough in-band data in the input buffer that normal flow control prevents the peer from sending the urgent data until the buffer is cleared. The process must then read enough of the queued data before the urgent data can be delivered.

CODE EXAMPLE 2–12 Flushing Terminal I/O on Receipt of Out-of-Band Data

```
#include <sys/ioctl.h>
#include <sys/file.h>
...
oob()
{
  int out = FWRITE;
  char waste[BUFSIZ];
  int mark = 0;

  /* flush local terminal output */
  ioctl(1, TIOCFLUSH, (char *) &out);
  while(1) {
   if (ioctl(rem, SIOCATMARK, &mark) == -1) {
    perror("ioctl");
    break;
   }
   if (mark)
    break;
   (void) read(rem, waste, sizeof waste);
```

```
  }
  if (recv(rem, &mark, 1, MSG_OOB) == –1) {
   perror("recv");
   ...
  }
  ...
}
```

There is also a facility to retain the position of urgent in-line data in the socket stream. This is available as a socket-level option, SO_OOBINLINE. See the getsockopt(3SOCKET) manpage for usage. With this option, the position of urgent data (the mark) is retained, but the urgent data immediately follows the mark in the normal data stream returned without the MSG_OOB flag. Reception of multiple urgent indications causes the mark to move, but no out-of-band data are lost.

Nonblocking Sockets

Some applications require sockets that do not block. For example, requests that cannot complete immediately and would cause the process to be suspended (awaiting completion) are not executed. An error code would be returned. After a socket is created and any connection to another socket is made, it can be made nonblocking by issuing a fcntl(2) call, as shown in Code Example 2–13.

CODE EXAMPLE 2–13 Set Nonblocking Socket

```
#include <fcntl.h>
#include <sys/file.h>
...
int fileflags;
int s;
...
s = socket(AF_INET6, SOCK_STREAM, 0);
...
if (fileflags = fcntl(s, F_GETFL, 0) == –1)
  perror("fcntl F_GETFL");
  exit(1);
}
if (fcntl(s, F_SETFL, fileflags | FNDELAY) == –1)
  perror("fcntl F_SETFL, FNDELAY");
  exit(1);
}
...
```

When doing I/O on a nonblocking socket, check for the error EWOULDBLOCK (in errno.h), which occurs when an operation would normally block. accept(3SOCKET), connect(3SOCKET), send(3SOCKET), recv(3SOCKET), read(2), and write(2) can all return EWOULDBLOCK. If an operation such as a send(3SOCKET) cannot be done in its entirety, but partial writes work (such as when using a stream socket), the data that can be sent immediately are processed, and the return value is the amount actually sent.

Asynchronous Socket I/O

Asynchronous communication between processes is required in applications that handle multiple requests simultaneously. Asynchronous sockets must be SOCK_STREAM type. To make a socket asynchronous, you issue a fcntl(2) call, as shown in Code Example 2–14.

CODE EXAMPLE 2–14 Making a Socket Asynchronous

```
#include <fcntl.h>
#include <sys/file.h>
...
int fileflags;
int s;
...
s = socket(AF_INET6, SOCK_STREAM, 0);
...
if (fileflags = fcntl(s, F_GETFL ) == –1)
  perror("fcntl F_GETFL");
  exit(1);
}
if (fcntl(s, F_SETFL, fileflags | FNDELAY | FASYNC) == –1)
  perror("fcntl F_SETFL, FNDELAY | FASYNC");
  exit(1);
}
...
```

After sockets are initialized, connected, and made asynchronous, communication is similar to reading and writing a file asynchronously. A send(3SOCKET), write(2), recv(3SOCKET), or read(2) initiates a data transfer. A data transfer is completed by a signal-driven I/O routine, described in the next section.

Interrupt-Driven Socket I/O

The SIGIO signal notifies a process when a socket (actually any file descriptor) has finished a data transfer. The steps in using SIGIO are:

- Set up a SIGIO signal handler with the signal(3C) or sigvec(3UCB) calls.
- Use fcntl(2) to set the process ID or process group ID to route the signal to its own process ID or process group ID (the default process group of a socket is group 0).
- Convert the socket to asynchronous, as shown in “Asynchronous Socket I/O” on page 47.

Code Example 2–15 shows some sample code to allow a given process to receive information on pending requests as they occur for a socket. With the addition of a handler for SIGURG, this code can also be used to prepare for receipt of SIGURG signals.

CODE EXAMPLE 2–15 Asynchronous Notification of I/O Requests

```
#include <fcntl.h>
#include <sys/file.h>
 ...
signal(SIGIO, io_handler);
/* Set the process receiving SIGIO/SIGURG signals to us. */
if (fcntl(s, F_SETOWN, getpid()) < 0) {
  perror("fcntl F_SETOWN");
  exit(1);
}
```

Signals and Process Group ID

For SIGURG and SIGIO, each socket has a process number and a process group ID. These values are initialized to zero, but can be redefined at a later time with the F_SETOWN fcntl(2), as in the previous example. A positive third argument to fcntl(2) sets the socket's process ID. A negative third argument to fcntl(2) sets the socket's process group ID. The only allowed recipient of SIGURG and SIGIO signals is the calling process. A similar fcntl(2), F_GETOWN, returns the process number of a socket.

Reception of SIGURG and SIGIO can also be enabled by using ioctl(2) to assign the socket to the user's process group:

```
/* oobdata is the out-of-band data handling routine */
sigset(SIGURG, oobdata);
int pid = -getpid();
if (ioctl(client, SIOCSPGRP, (char *) &pid) < 0) {
  perror("ioctl: SIOCSPGRP");
}
```

Another signal that is useful in server processes is SIGCHLD. This signal is delivered to a process when any child process changes state. Normally, servers use the signal to "reap" child processes that have exited without explicitly awaiting their termination or periodically polling for exit status. For example, the remote login server loop shown previously can be augmented as shown in Code Example 2–16.

CODE EXAMPLE 2–16 SIGCHLD Signal

```
int reaper();
...
sigset(SIGCHLD, reaper);
listen(f, 5);
while (1) {
  int g, len = sizeof from;
  g = accept(f, (struct sockaddr *) &from, &len);
  if (g < 0) {
   if (errno != EINTR)
    syslog(LOG_ERR, "rlogind: accept: %m");
   continue;
  }
  ...
}
```

```
#include <wait.h>

reaper()
{
  int options;
  int error;
  siginfo_t info;

  options = WNOHANG | WEXITED;
  bzero((char *) &info, sizeof(info));
  error = waitid(P_ALL, 0, &info, options);
}
```

If the parent server process fails to reap its children, zombie processes result.

Selecting Specific Protocols

If the third argument of the `socket`(3SOCKET) call is `0`, `socket`(3SOCKET) selects a default protocol to use with the returned socket of the type requested. The default protocol is usually correct, and alternate choices are not usually available. When using “raw” sockets to communicate directly with lower-level protocols or hardware interfaces, it can be important for the protocol argument to set up de-multiplexing. For example, raw sockets in the Internet family can be used to implement a new protocol on IP, and the socket receives packets only for the protocol specified. To obtain a particular protocol, determine the protocol number as defined in the protocol family. For the Internet family, use one of the library routines discussed in “Standard Routines” on page 34, such as `getprotobyname`(3SOCKET):

```
#include <sys/types.h>
#include <sys/socket.h>
#include <netinet/in.h>
#include <netdb.h>
 ...
pp = getprotobyname("newtcp");
s = socket(AF_INET6, SOCK_STREAM, pp->p_proto);
```

This results in a socket `s` using a stream-based connection, but with protocol type of `newtcp` instead of the default `tcp`.

Address Binding

TCP and UDP use a 4-tuple of *local IP address*, *local port number*, *foreign IP address*, and *foreign port number* to do their addressing. TCP requires these 4-tuples to be unique. UDP does not. It is unrealistic to expect user programs to always know proper values to use for the local address and local port, since a host can reside on multiple networks and the set of allocated port numbers is not directly accessible to a user. To avoid these problems, you can leave parts of the address unspecified and

let the system assign the parts appropriately when needed. Various portions of these tuples may be specified by various parts of the sockets API.

`bind`**(3SOCKET)**	Local address or local port or both
`connect`**(3SOCKET)**	Foreign address and foreign port

A call to `accept`(3SOCKET) retrieves connection information from a foreign client, so it causes the local address and port to be specified to the system (even though the caller of `accept`(3SOCKET) didn't specify anything), and the foreign address and port to be returned.

A call to `listen`(3SOCKET) can cause a local port to be chosen. If no explicit `bind`(3SOCKET) has been done to assign local information, `listen`(3SOCKET) causes an ephemeral port number to be assigned.

A service that resides at a particular port, but which does not care what local address is chosen, can `bind`(3SOCKET) itself to its port and leave the local address unspecified (set to `in6addr_any`, a variable with a constant value in `<netinet/in.h>`). If the local port need not be fixed, a call to `listen`(3SOCKET) causes a port to be chosen. Specifying an address of `in6addr_any` or a port number of 0 is known as wildcarding. (For `AF_INET`, `INADDR_ANY` is used in place of `in6addr_any`.)

The wildcard address simplifies local address binding in the Internet family. The sample code below binds a specific port number, MYPORT, to a socket, and leaves the local address unspecified.

CODE EXAMPLE 2–17 Bind Port Number to Socket

```
#include <sys/types.h>
#include <netinet/in.h>
...
struct sockaddr_in6 sin;
...
  s = socket(AF_INET6, SOCK_STREAM, 0);
  bzero (&sin6, sizeof (sin6));
  sin.sin6_family = AF_INET6;
  sin.sin6_addr.s6_addr = in6addr_any;
  sin.sin6_port = htons(MYPORT);
  bind(s, (struct sockaddr *) &sin, sizeof sin);
```

Each network interface on a host typically has a unique IP address. Sockets with wildcard local addresses can receive messages directed to the specified port number and sent to any of the possible addresses assigned to a host. For example, if a host has two interfaces with addresses 128.32.0.4 and 10.0.0.78, and a socket is bound as in Code Example 2–17, the process can accept connection requests addressed to 128.32.0.4 or 10.0.0.78. To allow only hosts on a specific network to connect to it, a server binds the address of the interface on the appropriate network.

Similarly, a local port number can be left unspecified (specified as `0`), in which case the system selects a port number. For example, to bind a specific local address to a socket, but to leave the local port number unspecified:

```
bzero (&sin, sizeof (sin));
(void) inet_pton (AF_INET6, ":ffff:127.0.0.1", sin.sin6_addr.s6_addr);
sin.sin6_family = AF_INET6;
sin.sin6_port = htons(0);
bind(s, (struct sockaddr *) &sin, sizeof sin);
```

The system uses two criteria to select the local port number:

- The first is that Internet port numbers less than 1024 (IPPORT_RESERVED) are reserved for privileged users (that is, the superuser). Nonprivileged users can use any Internet port number greater than 1024. The largest Internet port number is 65535.
- The second criterion is that the port number is not currently bound to some other socket.

The port number and IP address of the client is found through either accept(3SOCKET) (the *from* result) or getpeername(3SOCKET).

In certain cases, the algorithm used by the system to select port numbers is unsuitable for an application. This is because associations are created in a two-step process. For example, the Internet file transfer protocol specifies that data connections must always originate from the same local port. However, duplicate associations are avoided by connecting to different foreign ports. In this situation, the system would disallow binding the same local address and port number to a socket if a previous data connection's socket still existed. To override the default port selection algorithm, you must perform an option call before address binding:

```
 ...
int on = 1;
...
setsockopt(s, SOL_SOCKET, SO_REUSEADDR, &on, sizeof on);
bind(s, (struct sockaddr *) &sin, sizeof sin);
```

With this call, local addresses already in use can be bound. This does not violate the uniqueness requirement, because the system still verifies at connect time that any other sockets with the same local address and port do not have the same foreign address and port. If the association already exists, the error EADDRINUSE is returned.

Using Multicast

IP multicasting is only supported on AF_INET6 and AF_INETsockets of type SOCK_DGRAM and SOCK_RAW, and only on subnetworks for which the interface driver supports multicasting.

Sending IPv4 Multicast Datagrams

To send a multicast datagram, specify an IP multicast address in the range 224.0.0.0 to 239.255.255.255 as the destination address in a sendto(3SOCKET) call.

By default, IP multicast datagrams are sent with a time-to-live (TTL) of 1, which prevents them from being forwarded beyond a single subnetwork. The socket option `IP_MULTICAST_TTL` allows the TTL for subsequent multicast datagrams to be set to any value from 0 to 255, to control the scope of the multicasts:

```
u_char ttl;
setsockopt(sock, IPPROTO_IP, IP_MULTICAST_TTL, &ttl,sizeof(ttl))
```

Multicast datagrams with a TTL of 0 are not transmitted on any subnet, but can be delivered locally if the sending host belongs to the destination group and if multicast loopback has not been disabled on the sending socket (see below). Multicast datagrams with TTL greater than one can be delivered to more than one subnet if one or more multicast routers are attached to the first-hop subnet. To provide meaningful scope control, the multicast routers support the notion of TTL "thresholds", which prevent datagrams with less than a certain TTL from traversing certain subnets. The thresholds enforce the following convention that multicast datagrams with initial TTL:

0	Are restricted to the same host
1	Are restricted to the same subnet
32	Are restricted to the same site
64	Are restricted to the same region
128	Are restricted to the same continent
255	Are unrestricted in scope

"Sites" and "regions" are not strictly defined, and sites can be subdivided into smaller administrative units, as a local matter.

An application can choose an initial TTL other than the ones listed above. For example, an application might perform an "expanding-ring search" for a network resource by sending a multicast query, first with a TTL of 0, and then with larger and larger TTLs, until a reply is received, using (for example) the TTL sequence 0, 1, 2, 4, 8, 16, 32.

The multicast router refuses to forward any multicast datagram with a destination address between 224.0.0.0 and 224.0.0.255, inclusive, regardless of its TTL. This range of addresses is reserved for the use of routing protocols and other low-level topology discovery or maintenance protocols, such as gateway discovery and group membership reporting.

Each multicast transmission is sent from a single network interface, even if the host has more than one multicast-capable interface. (If the host is also a multicast router

and the TTL is greater than 1, a multicast can be *forwarded* to interfaces other than originating interface.) A socket option is available to override the default for subsequent transmissions from a given socket:

```
struct in_addr addr;
setsockopt(sock, IPPROTO_IP, IP_MULTICAST_IF, &addr, sizeof(addr))
```

where `addr` is the local IP address of the outgoing interface you want. Revert to the default interface by specifying the address `INADDR_ANY`. The local IP address of an interface is obtained with the `SIOCGIFCONF` ioctl. To determine if an interface supports multicasting, fetch the interface flags with the `SIOCGIFFLAGS` ioctl and test if the `IFF_MULTICAST` flag is set. (This option is intended primarily for multicast routers and other system services specifically concerned with internet topology.)

If a multicast datagram is sent to a group to which the sending host itself belongs (on the outgoing interface), a copy of the datagram is, by default, looped back by the IP layer for local delivery. Another socket option gives the sender explicit control over whether or not subsequent datagrams are looped back:

```
u_char loop;
setsockopt(sock, IPPROTO_IP, IP_MULTICAST_LOOP, &loop, sizeof(loop))
```

where `loop` is 0 to disable loopback, and 1 to enable loopback. This option provides a performance benefit for applications that have only one instance on a single host (such as a router or a mail demon), by eliminating the overhead of receiving their own transmissions. It should not normally be used by applications that can have more than one instance on a single host (such as a conferencing program) or for which the sender does not belong to the destination group (such as a time querying program).

If the sending host belongs to the destination group on another interface, a multicast datagram sent with an initial TTL greater than 1 can be delivered to the sending host on the other interface. The loopback control option has no effect on such delivery.

Receiving IPv4 Multicast Datagrams

Before a host can receive IP multicast datagrams, it must become a member of one, or more, IP multicast group. A process can ask the host to join a multicast group by using the following socket option:

```
struct ip_mreq mreq;
setsockopt(sock, IPPROTO_IP, IP_ADD_MEMBERSHIP, &mreq, sizeof(mreq))
```

where `mreq` is the structure

```
struct ip_mreq {
    struct in_addr imr_multiaddr;   /* multicast group to join */
```

```
        struct in_addr imr_interface;   /* interface to join on */
    }
```

Each membership is associated with a single interface, and it is possible to join the same group on more than one interface. Specify `imr_interface` to be `in6addr_any` to choose the default multicast interface, or one of the host's local addresses to choose a particular (multicast-capable) interface.

To drop a membership, use

```
struct ip_mreq mreq;
setsockopt(sock, IPPROTO_IP, IP_DROP_MEMBERSHIP, &mreq, sizeof(mreq))
```

where `mreq` contains the same values used to add the membership. The memberships associated with a socket are also dropped when the socket is closed or the process holding the socket is killed. More than one socket can claim a membership in a particular group, and the host remains a member of that group until the last claim is dropped.

Incoming multicast packets are accepted by the kernel IP layer if any socket has claimed a membership in the destination group of the datagram. Delivery of a multicast datagram to a particular socket is based on the destination port and the memberships associated with the socket (or protocol type, for raw sockets), just as with unicast datagrams. To receive multicast datagrams sent to a particular port, bind to the local port, leaving the local address unspecified (such as, `INADDR_ANY`).

More than one process can bind to the same `SOCK_DGRAM` UDP port if the `bind`(3SOCKET) is preceded by:

```
int one = 1;
setsockopt(sock, SOL_SOCKET, SO_REUSEADDR, &one, sizeof(one))
```

In this case, every incoming multicast or broadcast UDP datagram destined to the shared port is delivered to all sockets bound to the port. For backwards compatibility reasons, *this does not apply to incoming unicast datagrams.* Unicast datagrams are never delivered to more than one socket, regardless of how many sockets are bound to the datagram's destination port. `SOCK_RAW` sockets do not require the `SO_REUSEADDR` option to share a single IP protocol type.

The definitions required for the new, multicast-related socket options are found in `<netinet/in.h>`. All IP addresses are passed in network byte-order.

Sending IPv6 Multicast Datagrams

To send a multicast datagram, specify an IP multicast address in the range `ff00::0/8` as the destination address in a `sendto`(3SOCKET) call.

By default, IP multicast datagrams are sent with a hop limit of 1, which prevents them from being forwarded beyond a single subnetwork. The socket option

IPV6_MULTICAST_HOPS allows the hoplimit for subsequent multicast datagrams to be set to any value from 0 to 255, to control the scope of the multicasts:

```
uint_l;
setsockopt(sock, IPPROTO_IPV6, IPV6_MULTICAST_HOPS, &hops,sizeof(hops))
```

Multicast datagrams with a hoplimit of 0 are not transmitted on any subnet, but can be delivered locally if the sending host belongs to the destination group and if multicast loopback has not been disabled on the sending socket (see below). Multicast datagrams with hoplimit greater than one can be delivered to more than one subnet if one or more multicast routers are attached to the first-hop subnet. The IPv6 multicast addresses, unlike their IPv4 counterparts, contain explicit scope information encoded in the first part of the address. The defined scopes are (where X is unspecified):

ffX1::0/16	Node–local scope — restricted to the same node
ffX2::0/16	Link–local scope
ffX5::0/16	Site–local scope
ffX8::0/16	Organization–local scope
ffXe::0/16	Global scope

An application can, separately from the scope of the multicast address, use different hoplimit values. For example, an application might perform an "expanding-ring search" for a network resource by sending a multicast query, first with a hoplimit of 0, and then with larger and larger hoplimits, until a reply is received, using (for example) the hoplimit sequence 0, 1, 2, 4, 8, 16, 32.

Each multicast transmission is sent from a single network interface, even if the host has more than one multicast-capable interface. (If the host is also a multicast router and the hoplimit is greater than 1, a multicast can be *forwarded* to interfaces other than originating interface.) A socket option is available to override the default for subsequent transmissions from a given socket:

```
uint_t ifindex;

ifindex = if_nametoindex )"hme3");
setsockopt(sock, IPPROTO_IPV6, IPV6_MULTICAST_IF, &ifindex, sizeof(ifindex))
```

where ifindex is the interface index for the desired outgoing interface. Revert to the default interface by specifying the value 0.

If a multicast datagram is sent to a group to which the sending host itself belongs (on the outgoing interface), a copy of the datagram is, by default, looped back by the

IP layer for local delivery. Another socket option gives the sender explicit control over whether or not subsequent datagrams are looped back:

```
uint_t loop;
setsockopt(sock, IPPROTO_IPV6, IPV6_MULTICAST_LOOP, &loop, sizeof(loop))
```

where `loop` is 0 to disable loopback, and 1 to enable loopback. This option provides a performance benefit for applications that have only one instance on a single host (such as a router or a mail demon), by eliminating the overhead of receiving their own transmissions. It should not normally be used by applications that can have more than one instance on a single host (such as a conferencing program) or for which the sender does not belong to the destination group (such as a time querying program).

If the sending host belongs to the destination group on another interface, a multicast datagram sent with an initial hoplimit greater than 1 can be delivered to the sending host on the other interface. The loopback control option has no effect on such delivery.

Receiving IPv6 Multicast Datagrams

Before a host can receive IP multicast datagrams, it must become a member of one, or more, IP multicast group. A process can ask the host to join a multicast group by using the following socket option:

```
struct ipv6_mreq mreq;
setsockopt(sock, IPPROTO_IPV6, IPV6_JOIN_GROUP, &mreq, sizeof(mreq))
```

where `mreq` is the structure

```
struct ipv6_mreq {
    struct in6_addr ipv6mr_multiaddr;   /* IPv6 multicast addr */
    unsigned int    ipv6mr_interface;   /* interface index */
}
```

Each membership is associated with a single interface, and it is possible to join the same group on more than one interface. Specify `ipv6_interface` to be `0` to choose the default multicast interface, or an interface index for one of the host's interfaces to choose that (multicast capable) interface.

To leave a group, use

```
struct ipv6_mreq mreq;
setsockopt(sock, IPPROTO_IPV6, IP_LEAVE_GROUP, &mreq, sizeof(mreq))
```

where `mreq` contains the same values used to add the membership. The memberships associated with a socket are also dropped when the socket is closed or the process holding the socket is killed. More than one socket can claim a

membership in a particular group, and the host remains a member of that group until the last claim is dropped.

Incoming multicast packets are accepted by the kernel IP layer if any socket has claimed a membership in the destination group of the datagram. Delivery of a multicast datagram to a particular socket is based on the destination port and the memberships associated with the socket (or protocol type, for raw sockets), just as with unicast datagrams. To receive multicast datagrams sent to a particular port, bind to the local port, leaving the local address unspecified (such as, `INADDR_ANY`).

More than one process can bind to the same `SOCK_DGRAM` UDP port if the `bind`(3SOCKET) is preceded by:

```
int one = 1;
setsockopt(sock, SOL_SOCKET, SO_REUSEADDR, &one, sizeof(one))
```

In this case, every incoming multicast UDP datagram destined to the shared port is delivered to all sockets bound to the port. For backwards compatibility reasons, *this does not apply to incoming unicast datagrams*. Unicast datagrams are never delivered to more than one socket, regardless of how many sockets are bound to the datagram's destination port. `SOCK_RAW` sockets do not require the `SO_REUSEADDR` option to share a single IP protocol type.

The definitions required for the new, multicast-related socket options are found in `<netinet/in.h>`. All IP addresses are passed in network byte-order.

Zero Copy and Checksum Off-load

In SunOS 5.6 and later, the TCP/IP protocol stack has been enhanced to support two new features: zero copy and TCP checksum off-load.

- Zero copy uses virtual memory MMU remapping and a copy-on-write technique to move data between the application and the kernel space.
- Checksum off-loading relies on special hardware logic to off-load the TCP checksum calculation.

Note - Although zero copy and checksum off-loading are functionally independent of one another, they have to work together to obtain the optimal performance. Checksum off-loading requires hardware support from the network interface and, without this hardware support, zero copy is not enabled.

Zero copy requires that the applications supply page-aligned buffers before VM page remapping can be applied. Applications should use large, circular buffers on the transmit side to avoid expensive copy-on-write faults. A typical buffer allocation is sixteen 8K buffers.

Socket Options

You can set and get several options on sockets through `setsockopt`(3SOCKET) and `getsockopt`(3SOCKET); for example by changing the send or receive buffer space. The general forms of the calls are:

```
setsockopt(s, level, optname, optval, optlen);
```

and

```
getsockopt(s, level, optname, optval, optlen);
```

In some cases, such as setting the buffer sizes, these are only hints to the operating system. The operating system reserves the right to adjust the values appropriately.

Table 2–4 shows the arguments of the calls.

TABLE 2–4 `setsockopt`(3SOCKET) and `getsockopt`(3SOCKET) Arguments

Arguments	Description
s	Socket on which the option is to be applied
level	Specifies the protocol level, such as socket level, indicated by the symbolic constant `SOL_SOCKET` in `sys/socket.h`
optname	Symbolic constant defined in `sys/socket.h` that specifies the option
optval	Points to the value of the option
optlen	Points to the length of the value of the option

For `getsockopt`(3SOCKET), *optlen* is a value-result argument, initially set to the size of the storage area pointed to by *optval* and set on return to the length of storage used.

It is sometimes useful to determine the type (for example, stream or datagram) of an existing socket. Programs invoked by `inetd`(1M) can do this by using the `SO_TYPE` socket option and the `getsockopt`(3SOCKET) call:

```
#include <sys/types.h>
#include <sys/socket.h>

int type, size;

size = sizeof (int);
```

```
if (getsockopt(s, SOL_SOCKET, SO_TYPE, (char *) &type, &size) <0) {
  ...
}
```

After `getsockopt`(3SOCKET), `type` is set to the value of the socket type, as defined in `sys/socket.h`. For a datagram socket, `type` would be `SOCK_DGRAM`.

`inetd`(1M) Daemon

One of the daemons provided with the system is `inetd`(1M). It is invoked at start-up time, and gets the services for which it listens from the `/etc/inet/inetd.conf` file. The daemon creates one socket for each service listed in `/etc/inet/inetd.conf`, binding the appropriate port number to each socket. See the `inetd`(1M) man page for details.

`inetd`(1M) polls each socket, waiting for a connection request to the service corresponding to that socket. For `SOCK_STREAM` type sockets, `inetd`(1M) does an `accept`(3SOCKET) on the listening socket, `fork`(2)s, `dup`(2)s the new socket to file descriptors `0` and `1` (`stdin` and `stdout`), closes other open file descriptors, and `exec`(2)s the appropriate server.

The primary benefit of `inetd`(1M) is that services that are not in use are not taking up machine resources. A secondary benefit is that `inetd`(1M) does most of the work to establish a connection. The server started by `inetd`(1M) has the socket connected to its client on file descriptors `0` and `1`, and can immediately `read`(2), `write`(2), `send`(3SOCKET), or `recv`(3SOCKET). Servers can use buffered I/O as provided by the `stdio` conventions, as long as they use `fflush`(3C) when appropriate.

`getpeername`(3SOCKET) returns the address of the peer (process) connected to a socket; it is useful in servers started by `inetd`(1M). For example, to log the Internet address (such as `fec0::56:a00:20ff:fe7d:3dd2`, which is conventional for representing the IPv6 address of a client), an `inetd`(1M) server could use the following:

```
struct sockaddr_storage name;
int namelen = sizeof (name);
char abuf[INET6_ADDRSTRLEN];
struct in6_addr addr6;
struct in_addr addr;

if (getpeername(fd, (struct sockaddr *)&name, &namelen) == -1) {
    perror("getpeername");
    exit(1);
} else {
    addr = ((struct sockaddr_in *)&name)->sin_addr;
    addr6 = ((struct sockaddr_in6 *)&name)->sin6_addr;
    if (name.ss_family == AF_INET) {
            (void) inet_ntop(AF_INET, &addr, abuf, sizeof (abuf));
    } else if (name.ss_family == AF_INET6 && IN6_IS_ADDR_V4MAPPED(&addr6)) {
            /* this is a IPv4-mapped IPv6 address */
            IN6_MAPPED_TO_IN(&addr6, &addr);
            (void) inet_ntop(AF_INET, &addr, abuf, sizeof (abuf));
```

```
        } else if (name.ss_family == AF_INET6) {
                (void) inet_ntop(AF_INET6, &addr6, abuf, sizeof (abuf));

        }
        syslog("Connection from %s\n", abuf);
    }
```

Broadcasting and Determining Network Configuration

Broadcasting is not supported in IPv6. It is supported only in IPv4.

Messages sent by datagram sockets can be broadcast to reach all of the hosts on an attached network. The network must support broadcast; the system provides no simulation of broadcast in software. Broadcast messages can place a high load on a network since they force every host on the network to service them. Broadcasting is usually used for either of two reasons: to find a resource on a local network without having its address, or functions like routing require that information be sent to all accessible neighbors.

To send a broadcast message, create an Internet datagram socket:

```
s = socket(AF_INET, SOCK_DGRAM, 0);
```

and bind a port number to the socket:

```
sin.sin_family = AF_INET;
sin.sin_addr.s_addr = htonl(INADDR_ANY);
sin.sin_port = htons(MYPORT);
bind(s, (struct sockaddr *) &sin, sizeof sin);
```

The datagram can be broadcast on only one network by sending to the network's broadcast address. A datagram can also be broadcast on all attached networks by sending to the special address `INADDR_BROADCAST`, defined in `netinet/in.h`.

The system provides a mechanism to determine a number of pieces of information (including the IP address and broadcast address) about the network interfaces on the system. The `SIOCGIFCONF ioctl`(2) call returns the interface configuration of a host in a single `ifconf` structure. This structure contains an array of `ifreq` structures, one for each address family supported by each network interface to which the host is connected. Code Example 2–18 shows these structures defined in `net/if.h`.

CODE EXAMPLE 2–18 `net/if.h` Header File

```
struct ifreq {
#define IFNAMSIZ 16
char ifr_name[IFNAMSIZ]; /* if name, e.g., "en0" */
union {
  struct sockaddr ifru_addr;
  struct sockaddr ifru_dstaddr;
```

```
  char ifru_oname[IFNAMSIZ]; /* other if name */
  struct sockaddr ifru_broadaddr;
  short ifru_flags;
  int ifru_metric;
  char ifru_data[1]; /* interface dependent data */
  char ifru_enaddr[6];
} ifr_ifru;
#define ifr_addr ifr_ifru.ifru_addr
#define ifr_dstaddr ifr_ifru.ifru_dstaddr
#define ifr_oname ifr_ifru.ifru_oname
#define ifr_broadaddr ifr_ifru.ifru_broadaddr
#define ifr_flags ifr_ifru.ifru_flags
#define ifr_metric ifr_ifru.ifru_metric
#define ifr_data ifr_ifru.ifru_data
#define ifr_enaddr ifr_ifru.ifru_enaddr
};
```

The call that obtains the interface configuration is:

```
/*
 * Do SIOCGIFNUM ioctl to find the number of interfaces
 *
 * Allocate space for number of interfaces found
 *
 * Do SIOCGIFCONF with allocated buffer
 *
 */
if (ioctl(s, SIOCGIFNUM, (char *)&numifs) == –1) {
        numifs = MAXIFS;
}
bufsize = numifs * sizeof(struct ifreq);
reqbuf = (struct ifreq *)malloc(bufsize);
if (reqbuf == NULL) {
        fprintf(stderr, "out of memory\n");
        exit(1);
}
ifc.ifc_buf = (caddr_t)&reqbuf[0];
ifc.ifc_len = bufsize;
if (ioctl(s, SIOCGIFCONF, (char *)&ifc) == –1) {
        perror("ioctl(SIOCGIFCONF)");
        exit(1);
}
...
}
```

After this call, *buf* contains an array of `ifreq` structures, one for each network to which the host is connected. These structures are ordered first by interface name, then by supported address families. `ifc.ifc_len` is set to the number of bytes used by the `ifreq` structures.

Each structure has a set of interface flags that tell whether the corresponding network is up or down, point-to-point or broadcast, and so on. Code Example 2–19 shows the `SIOCGIFFLAGS ioctl`(2) returning these flags for an interface specified by an `ifreq` structure.

CODE EXAMPLE 2–19 Obtaining Interface Flags

```
struct ifreq *ifr;
ifr = ifc.ifc_req;
for (n = ifc.ifc_len/sizeof (struct ifreq); —n >= 0; ifr++) {
   /*
    * Be careful not to use an interface devoted to an address
    * family other than those intended.
    */
   if (ifr->ifr_addr.sa_family != AF_INET)
      continue;
   if (ioctl(s, SIOCGIFFLAGS, (char *) ifr) < 0) {
      ...
   }
   /* Skip boring cases */
   if ((ifr->ifr_flags & IFF_UP) == 0 ||
      (ifr->ifr_flags & IFF_LOOPBACK) ||
      (ifr->ifr_flags & (IFF_BROADCAST | IFF_POINTOPOINT)) == 0)
      continue;
}
```

Code Example 2–20 shows the broadcast of an interface can be obtained with the `SIOGGIFBRDADDR ioctl`(2).

CODE EXAMPLE 2–20 Broadcast Address of an Interface

```
if (ioctl(s, SIOCGIFBRDADDR, (char *) ifr) < 0) {
  ...
}
memcpy((char *) &dst, (char *) &ifr->ifr_broadaddr,
  sizeof ifr->ifr_broadaddr);
```

The `SIOGGIFBRDADDR ioctl`(2) can also be used to get the destination address of a point-to-point interface.

After the interface broadcast address is obtained, transmit the broadcast datagram with `sendto`(3SOCKET):

```
sendto(s, buf, buflen, 0, (struct sockaddr *)&dst, sizeof dst);
```

Use one `sendto`(3SOCKET) for each interface to which the host is connected that supports the broadcast or point-to-point addressing.

CHAPTER 3

Programming With XTI and TLI

The X/Open Transport Interface (XTI) and the Transport Layer Interface (TLI) are a set of functions that constitute a network programming interface. XTI is an evolution from the older TLI interface available on SunOS 4. Both interfaces are supported, though XTI represents the future direction of this set of interfaces.

XTI/TLI Is Multithread Safe

The interfaces described in this chapter are multithread safe. This means that applications containing XTI/TLI function calls can be used freely in a multithreaded application. However, the degree of concurrency available to applications is not specified.

XTI/TLI Are Not Asynchronous Safe

The XTI/TLI interface behavior has not been well specified in an asynchronous environment. It is not recommended that these interfaces be used from signal handler routines.

What Are XTI and TLI?

TLI was introduced with AT&T's System V, Release 3 in 1986. It provided a transport layer interface API. TLI was modeled after the ISO Transport Service Definition and provides an API between the OSI transport and session layers. TLI interfaces evolved further in AT&T System V, Release 4 version of Unix and were made available in SunOS 5.6 operating system interfaces, too.

XTI interfaces are an evolution of TLI interfaces and represent the future direction of this family of interfaces. Compatibility for applications using TLI interfaces is available. There is no intrinsic need to port TLI applications to XTI immediately. New applications can use the XTI interfaces and older applications can be ported to XTI when necessary.

TLI is implemented as a set of function calls in a library (`libnsl`) with which the applications link. XTI applications are compiled using the c89 front end and must be linked with the `xnet` library (`libxnet`). For additional information on compiling with XTI, see `standards`(5).

Note - An application using the XTI interface uses the `xti.h` header file, whereas an application using the TLI interface includes the `tiuser.h` header file.

Intrinsic to XTI/TLI are the notions of *transport endpoints* and a *transport provider*. The transport endpoints are two entities that are communicating, and the transport provider is the set of routines on the host that provides the underlying communication support. XTI/TLI is the interface to the transport provider, not the provider itself. See Figure 3–1.

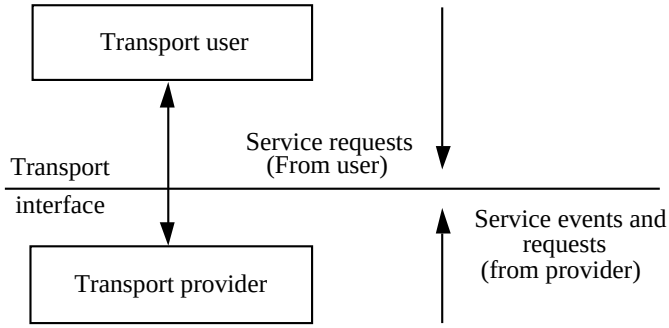

Figure 3–1 How XTI/TLI Works

XTI/TLI code can be written to be independent of current transport providers in conjunction with some additional interfaces and mechanisms described in Chapter 4. The SunOS 5 product includes some transport providers (TCP, for example) as part of the base operating system. A transport provider performs services, and the transport user requests the services. The transport user issues service requests to the transport provider. An example is a request to transfer data over a connection TCP and UDP.

XTI/TLI can also be used for transport-independent programming. XTI/TLI has two components to achieve this:

- Library routines that perform the transport services, in particular, transport selection and name-to-address translation. The network services library includes a set of functions that implement XTI/TLI for user processes. See Chapter 4.

  Programs using TLI should be linked with the network services library, `libnsl`, as follows:

```
% cc prog.c -lnsl
```

- State transition rules that define the sequence in which the transport routines can be invoked. For more information on state transition rules, see section, “State Transitions” on page 104. The state tables define the legal sequence of library calls based on the state and the handling of events. These events include user-generated library calls, as well as provider-generated event indications. XTI/TLI programmers should understand all state transitions before using the interface.

XTI/TLI provides two modes of service: connection mode and connectionless mode. The next two sections give an overview of these modes.

Connectionless Mode

Connectionless mode is message oriented. Data are transferred in self-contained units with no relationship between the units. This service requires only an established association between the peer users that determines the characteristics of the data. All information required to deliver a message (such as the destination address) is presented to the transport provider, with the data to be transmitted, in one service request. Each message is entirely self-contained. Use connectionless mode service for applications that:

- Have short-term request / response interactions
- Are dynamically reconfigurable
- Do not require sequential delivery of data

Connectionless transports can be unreliable. They need not necessarily maintain message sequence, and messages are sometimes lost.

Connectionless Mode Routines

Connectionless-mode transport service has two phases: local management and data transfer. The local management phase defines the same local operations as for the connection mode service.

The data transfer phase lets a user transfer data units (usually called datagrams) to the specified peer user. Each data unit must be accompanied by the transport address of the destination user. `t_sndudata`(3NSL) sends and `t_rcvvudata`(3NSL) receives messages. Table 3–1summarizes all routines for connectionless mode data transfer.

TABLE 3–1 Routines for Connectionless-Mode Data Transfer

Command	Description
`t_sndudata`	Sends a message to another user of the transport
`t_rcvudata`	Receives a message sent by another user of the transport
`t_rcvuderr`	Retrieves error information associated with a previously sent message

Connectionless Mode Service

Connectionless mode service is appropriate for short-term request/response interactions, such as transaction-processing applications. Data are transferred in self-contained units with no logical relationship required among multiple units.

Endpoint Initiation

Transport users must initialize XTI/TLI endpoints before transferring data. They must choose the appropriate connectionless service provider using `t_open`(3NSL) and establish its identity using `t_bind`(3NSL).

Use `t_optmgmt`(3NSL) to negotiate protocol options. Like connection mode service, each transport provider specifies the options, if any, it supports. Option negotiation is a protocol-specific activity. In Code Example 3–1, the server waits for incoming queries, and processes and responds to each query. The example also shows the definitions and initiation sequence of the server.

CODE EXAMPLE 3–1 CLTS Server

```
#include <stdio.h>
#include <fcntl.h>
#include <xti.h> /* TLI applications use <tiuser.h>  */
#define SRV_ADDR 2 /* server's well known address */

main()
{
   int fd;
   int flags;
   struct t_bind *bind;
   struct t_unitdata *ud;
   struct t_uderr *uderr;
   extern int t_errno;

 if ((fd = t_open("/dev/exmp", O_RDWR, (struct t_info *) NULL))
         == –1) {
      t_error("unable to open /dev/exmp");
      exit(1);
   }
  if ((bind = (struct t_bind *)t_alloc(fd, T_BIND, T_ADDR))
         == (struct t_bind *) NULL) {
      t_error("t_alloc of t_bind structure failed");
      exit(2);
   }
   bind->addr.len = sizeof(int);
   *(int *)bind->addr.buf = SRV_ADDR;
   bind->qlen = 0;
   if (t_bind(fd, bind, bind) == –1) {
      t_error("t_bind failed");
      exit(3);
```

```
}
/*
 * TLI interface applications need the following code which
 * is no longer needed for XTI interface applications.
 * ------------------------------------
 * Verify if the bound address correct?
 *
 * if (bind -> addr.len != sizeof(int) ||
 *      *(int *)bind->addr.buf != SRV_ADDR) {
 * fprintf(stderr, "t_bind bound wrong address\n");
 * exit(4);
 * }
 * --------------------------------------
 */
```

The server establishes a transport endpoint with the desired transport provider using `t_open`(3NSL). Each provider has an associated service type, so the user can choose a particular service by opening the appropriate transport provider file. This connectionless mode server ignores the characteristics of the provider returned by `t_open`(3NSL) by setting the third argument to `NULL`. The transaction server assumes the transport provider has the following characteristics:

- The transport address is an integer value that uniquely identifies each user.
- The transport provider supports the `T_CLTS` service type (connectionless transport service, or datagram).
- The transport provider does not require any protocol-specific options.

The connectionless server binds a transport address to the endpoint so that potential clients can access the server. A `t_bind` structure is allocated using `t_alloc`(3NSL) and the `buf` and `len` fields of the address are set accordingly.

One difference between a connection mode server and a connectionless mode server is that the `qlen` field of the `t_bind` structure is `0` for connectionless mode service. There are no connection requests to queue.

XTI/TLI interfaces define an inherent client-server relationship between two users while establishing a transport connection in the connection mode service. No such relationship exists in connectionless mode service.

TLI requires that the server check the bound address returned by `t_bind`(3NSL) to ensure that it is the same as the one supplied. `t_bind`(3NSL) can also bind the endpoint to a separate, free address if the one requested is busy.

Data Transfer

After a user has bound an address to the transport endpoint, datagrams can be sent or received over the endpoint. Each outgoing message carries the address of the destination user. XTI/TLI also lets you specify protocol options to the transfer of the data unit (for example, transit delay). Each transport provider defines the set of

options on a datagram. When the datagram is passed to the destination user, the associated protocol options can be passed, too.

Code Example 3–2 illustrates the data transfer phase of the connectionless mode server.

CODE EXAMPLE 3–2 Data Transfer Routine

```
 if ((ud = (struct t_unitdata *) t_alloc(fd, T_UNITDATA,T_ALL))
         == (struct t_unitdata *) NULL) {
      t_error("t_alloc of t_unitdata struct failed");
      exit(5);
   }
   if ((uderr = (struct t_uderr *) t_alloc(fd, T_UDERROR, T_ALL))
         == (struct t_uderr *) NULL) {
      t_error("t_alloc of t_uderr struct failed");
      exit(6);
   }
   while(1) {
      if (t_rcvudata(fd, ud, &flags) == –1) {
         if (t_errno == TLOOK) {
               /* Error on previously sent datagram */
               if(t_rcvuderr(fd, uderr) == –1) {
                  exit(7);
               }
            fprintf(stderr, "bad datagram, error=%d\n",
               uderr->error);
            continue;
         }
         t_error("t_rcvudata failed");
         exit(8);
      }
      /*
       * Query() processes the request and places the response in
       * ud->udata.buf, setting ud->udata.len
       */
      query(ud);
      if (t_sndudata(fd, ud) == –1) {
         t_error("t_sndudata failed");
         exit(9);
      }
   }
}

/* ARGS USED */
void
query(ud)
struct t_unitdate *ud;
{
   /* Merely a stub for simplicity */
}
```

To buffer datagrams, the server first allocates a `t_unitdata` structure, which has the following format:

```
struct t_unitdata {
  struct netbuf addr;
  struct netbuf opt;
```

```
  struct netbuf udata;
}
```

addr holds the source address of incoming datagrams and the destination address of outgoing datagrams. opt holds any protocol options on the datagram. udata holds the data. The addr, opt, and udata fields must all be allocated with buffers large enough to hold any possible incoming values. The T_ALL argument of t_alloc(3NSL) ensures this and sets the maxlen field of each netbuf structure accordingly. The provider does not support protocol options in this example, so maxlen is set to 0 in the opt netbufstructure. The server also allocates a t_uderr structure for datagram errors.

The transaction server loops forever, receiving queries, processing the queries, and responding to the clients. It first calls t_rcvudata(3NSL) to receive the next query. t_rcvudata(3NSL) blocks until a datagram arrives, and returns it.

The second argument of t_rcvudata(3NSL) identifies the t_unitdata structure in which to buffer the datagram.

The third argument, flags, points to an integer variable and can be set to T_MORE on return from t_rcvudata(3NSL) to indicate that the user's udata buffer is too small to store the full datagram.

If this happens, the next call to t_rcvudata(3NSL) retrieves the rest of the datagram. Because t_alloc(3NSL) allocates a udata buffer large enough to store the maximum size datagram, this transaction server does not have to check flags. This is true only of t_rcvudata(3NSL) and not of any other receive primitives.

When a datagram is received, the transaction server calls its query routine to process the request. This routine stores a response in the structure pointed to by ud, and sets ud-->udata.len to the number of bytes in the response. The source address returned by t_rcvudata(3NSL) in ud-->addr is the destination address for t_sndudata(3NSL). When the response is ready, t_sndudata(3NSL) is called to send the response to the client.

Datagram Errors

If the transport provider cannot process a datagram sent by t_sndudata(3NSL), it returns a unit data error event, T_UDERR, to the user. This event includes the destination address and options of the datagram, and a protocol-specific error value that identifies the error. Datagram errors are protocol specific.

Note - A unit data error event does not always indicate success or failure in delivering the datagram to the specified destination. Remember, connectionless service does not guarantee reliable delivery of data.

The transaction server is notified of an error when it tries to receive another datagram. In this case, t_rcvudata(3NSL) fails, setting t_errno to TLOOK. If

TLOOK is set, the only possible event is T_UDERR, so the server calls t_rcvudata(3NSL) to retrieve the event. The second argument of t_rcvuderr(3NSL) is the t_uderr structure that was allocated earlier. This structure is filled in by t_rcvuderr(3NSL) and has the following format:

```
struct t_uderr {
  struct netbuf addr;
  struct netbuf opt;
  t_scalar_t error;
}
```

where addr and opt identify the destination address and protocol options specified in the bad datagram, and error is a protocol-specific error code. The transaction server prints the error code, then continues.

Connection Mode

Connection mode is circuit oriented. Data are transmitted in sequence over an established connection. The mode also provides an identification procedure that avoids address resolution and transmission in the data transfer phase. Use this service for applications that require data-stream-oriented interactions. Connection mode transport service has four phases:

- Local management
- Connection establishment
- Data transfer
- Connection release

The local management phase defines local operations between a transport user and a transport provider, as shown in Figure 3–2. For example, a user must establish a channel of communication with the transport provider. Each channel between a transport user and transport provider is a unique endpoint of communication, and is called the transport endpoint. t_open(3NSL) lets a user choose a particular transport provider to supply the connection mode services, and establishes the transport endpoint.

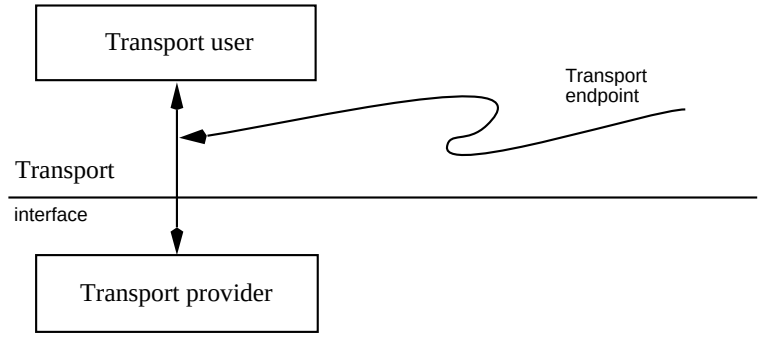

Figure 3–2 Transport Endpoint

Connection Mode Routines

Each user must establish an identity with the transport provider. A transport address is associated with each transport endpoint. One user process can manage several transport endpoints. In connection mode service, one user requests a connection to another user by specifying the other's address. The structure of a transport address is defined by the transport provider. An address can be as simple as an unstructured character string (for example, `file_server`), or as complex as an encoded bit pattern that specifies all information needed to route data through a network. Each transport provider defines its own mechanism for identifying users. Addresses can be assigned to the endpoint of a transport by `t_bind`(3NSL).

In addition to `t_open`(3NSL) and `t_bind`(3NSL), several routines support local operations. Table 3–2 summarizes all local management routines of XTI/TLI.

TABLE 3–2 Routines of XTI/TLI for Operating on the Endpoint

Command	Description
`t_alloc`	Allocates XTI/TLI data structures
`t_bind`	Binds a transport address to a transport endpoint
`t_close`	Closes a transport endpoint
`t_error`	Prints an XTI/TLI error message
`t_free`	Frees structures allocated using `t_alloc`(3NSL)

TABLE 3–2 Routines of XTI/TLI for Operating on the Endpoint *(continued)*

Command	Description
`t_getinfo`	Returns a set of parameters associated with a particular transport provider
`t_getprotaddr`	Returns the local and/or remote address associated with endpoint (XTI only)
`t_getstate`	Returns the state of a transport endpoint
`t_look`	Returns the current event on a transport endpoint
`t_open`	Establishes a transport endpoint connected to a chosen transport provider
`t_optmgmt`	Negotiates protocol-specific options with the transport provider
`t_sync`	Synchronizes a transport endpoint with the transport provider
`t_unbind`	Unbinds a transport address from a transport endpoint

The connection phase lets two users create a connection, or virtual circuit, between them, as shown in Figure 3–3.

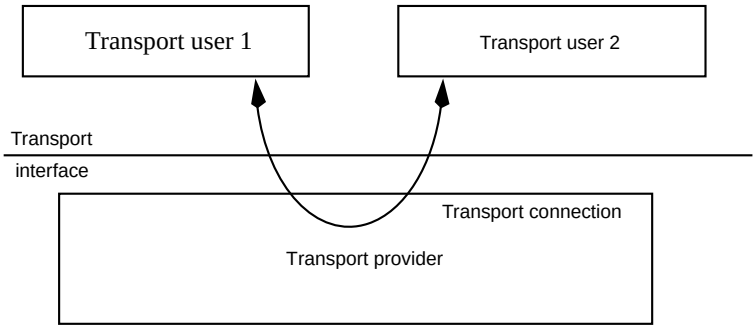

Figure 3–3 Transport Connection

For example, the connection phase occurs when a server advertises its service to a group of clients, then blocks on `t_listen`(3NSL) to wait for a request. A client tries to connect to the server at the advertised address by a call to `t_connect`(3NSL).

The connection request causes `t_listen`(3NSL) to return to the server, which can call `t_accept`(3NSL) to complete the connection.

Table 3–3 summarizes all routines available for establishing a transport connection. Refer to man pages for the specifications on these routines.

TABLE 3–3 Routines for Establishing a Transport Connection

Command	Description
`t_accept`	Accepts a request for a transport connection
`t_connect`	Establishes a connection with the transport user at a specified destination
`t_listen`	Listens for connect request from another transport user
`t_rcvconnect`	Completes connection establishment if `t_connect`(3NSL) was called in asynchronous mode (see "Advanced Topics" on page 93)

The data transfer phase lets users transfer data in both directions through the connection. `t_snd`(3NSL) sends and `t_rcv`(3NSL) receives data through the connection. It is assumed that all data sent by one user is guaranteed to be delivered to the other user in the order in which it was sent. Table 3–4 summarizes the connection mode data-transfer routines.

TABLE 3–4 Connection Mode Data Transfer Routines

Command	Description
`t_rcv`(3NSL)	Receives data that has arrived over a transport connection
`t_snd`(3NSL)	Sends data over an established transport connection

XTI/TLI has two types of connection release. The abortive release directs the transport provider to release the connection immediately. Any previously sent data that has not yet been transmitted to the other user can be discarded by the transport provider. `t_snddis`(3NSL) initiates the abortive disconnect. `t_rcvdis`(3NSL) receives the abortive disconnect. Transport providers usually support some form of abortive release procedure.

Some transport providers also support an orderly release that terminates communication without discarding data. `t_sndrel`(3NSL) and `t_rcvrel`(3NSL) perform this function. Table 3–5 summarizes the connection release routines. Refer to man pages for the specifications on these routines.

TABLE 3–5 Connection Release Routines

Command	Description
`t_rcvdis`(3NSL)	Returns a reason code for a disconnection and any remaining user data
`t_rcvrel`(3NSL)	Acknowledges receipt of an orderly release of a connection request
`t_snddis`(3NSL)	Aborts a connection or rejects a connect request
`t_sndrel`(3NSL)	Requests the orderly release of a connection

Connection Mode Service

The main concepts of connection mode service are illustrated through a client program and its server. The examples are presented in segments.

In the examples, the client establishes a connection to a server process. The server transfers a file to the client. The client receives the file contents and writes them to standard output.

Endpoint Initiation

Before a client and server can connect, each must first open a local connection to the transport provider (the transport endpoint) through `t_open`(3NSL), and establish its identity (or address) through `t_bind`(3NSL).

Many protocols perform a subset of the services defined in XTI/TLI. Each transport provider has characteristics that determine the services it provides and limit the

services. Data defining the transport characteristics are returned by t_open(3NSL) in a t_info structure. Table 3–6 shows the fields in a t_info structure.

TABLE 3–6 t_info Structure

Field	Content
addr	Maximum size of a transport address
options	Maximum bytes of protocol-specific options that can be passed between the transport user and transport provider
tsdu	Maximum message size that can be transmitted in either connection mode or connectionless mode
etsdu	Maximum expedited data message size that can be sent over a transport connection
connect	Maximum number of bytes of user data that can be passed between users during connection establishment
discon	Maximum bytes of user data that can be passed between users during the abortive release of a connection
servtype	The type of service supported by the transport provider

The three service types defined by XTI/TLI are:

1. T_COTS — The transport provider supports connection mode service but does not provide the orderly release facility. Connection termination is abortive, and any data not already delivered is lost.
2. T_COTS_ORD — The transport provider supports connection mode service with the orderly release facility.
3. T_CLTS — The transport provider supports connectionless mode service.

Only one such service can be associated with the transport provider identified by t_open(3NSL).

t_open(3NSL) returns the default provider characteristics of a transport endpoint. Some characteristics can change after an endpoint has been opened. This happens with negotiated options (option negotiation is described later in this section). t_getinfo(3NSL) returns the current characteristics of a transport endpoint.

After a user establishes an endpoint with the chosen transport provider, the client and server must establish their identities. t_bind(3NSL) does this by binding a transport address to the transport endpoint. For servers, this routine informs the transport provider that the endpoint is used to listen for incoming connect requests.

t_optmgmt(3NSL) can be used during the local management phase. It lets a user negotiate the values of protocol options with the transport provider. Each transport protocol defines its own set of negotiable protocol options, such as quality-of-service parameters. Because the options are protocol-specific, only applications written for a specific protocol use this function.

Client

The local management requirements of the example client and server are used to discuss details of these facilities. Code Example 3–3 shows the definitions needed by the client program, followed by its necessary local management steps.

CODE EXAMPLE 3–3 Client Implementation of Open and Bind

```
#include <stdio.h>
#include <tiuser.h>
#include <fcntl.h>
#define SRV_ADDR 1          /* server's address */

main()
{
   int fd;
   int nbytes;
   int flags = 0;
   char buf[1024];
   struct t_call *sndcall;
   extern int t_errno;

   if ((fd = t_open("/dev/exmp", O_RDWR, (struct t_info *),NULL))
         == –1) {
      t_error("t_open failed");
      exit(1);
   }
   if (t_bind(fd, (struct t_bind *) NULL, (struct t_bind *) NULL)
         == –1) {
      t_error("t_bind failed");
      exit(2);
   }
```

The first argument of t_open(3NSL) is the path of a file system object that identifies the transport protocol. /dev/exmp is the example name of a special file that identifies a generic, connection-based transport protocol. The second argument, O_RDWR, specifies to open for both reading and writing. The third argument points to a t_info structure in which to return the service characteristics of the transport.

This data is useful to write protocol-independent software (see “Guidelines to Protocol Independence” on page 112). In this example, a NULL pointer is passed. For Code Example 3–3, the transport provider must have the following characteristics:

- The transport address is an integer value that uniquely identifies each user.
- The transport provider supports the T_COTS_ORD service type, since the example uses orderly release.

- The transport provider does not require protocol-specific options.

If the user needs a service other than T_COTS_ORD, another transport provider can be opened. An example of the T_CLTS service invocation is shown in the section “Read / Write Interface” on page 91.

t_open(3NSL) returns the transport endpoint file handle that is used by all subsequent XTI / TLI function calls. The identifier is a file descriptor from opening the transport protocol file. See open(2).

The client then calls t_bind(3NSL) to assign an address to the endpoint. The first argument of t_bind(3NSL) is the transport endpoint handle. The second argument points to a t_bind structure that describes the address to bind to the endpoint. The third argument points to a t_bind structure that describes the address that the provider has bound.

The address of a client is rarely important because no other process tries to access it. That is why the second and third arguments to t_bind(3NSL) are NULL. The second NULL argument directs the transport provider to choose an address for the user.

If t_open(3NSL) or t_bind(3NSL) fails, the program calls t_error(3NSL) to display an appropriate error message by stderr. The global integer t_error(3NSL) is assigned an error value. A set of error values is defined in tiuser.h.

t_error(3NSL) is analogous to perror(3C). If the transport function error is a system error, t_errno(3NSL) is set to TSYSERR, and errno is set to the appropriate value.

Server

The server example must also establish a transport endpoint at which to listen for connection requests. Code Example 3–4 shows the definitions and local management steps.

CODE EXAMPLE 3–4 Server Implementation of Open and Bind

```
#include <tiuser.h>
#include <stropts.h>
#include <fcntl.h>
#include <stdio.h>
#include <signal.h>

#define DISCONNECT –1
#define SRV_ADDR 1        /* server’s address */
int conn_fd;   /* connection established here */
extern int t_errno;

main()
{
   int listen_fd;       /* listening transport endpoint */
   struct t_bind *bind;
   struct t_call *call;
```

```
if ((listen_fd = t_open("/dev/exmp", O_RDWR,
   (struct t_info *) NULL)) == –1) {
   t_error("t_open failed for listen_fd");
   exit(1);
}
if ((bind = (struct t_bind *)t_alloc( listen_fd, T_BIND, T_ALL))
      == (struct t_bind *) NULL) {
   t_error("t_alloc of t_bind structure failed");
   exit(2);
}
bind->qlen = 1;

/*
 * Because it assumes the format of the provider's address,
 * this program is transport-dependent
 */
 bind->addr.len = sizeof(int);
*(int *) bind->addr.buf = SRV_ADDR;
if (t_bind (listen_fd, bind, bind) < 0 ) {
   t_error("t_bind failed for listen_fd");
   exit(3);
}

#if (!defined(_XOPEN_SOURCE) ||(_XOPEN_SOURCE_EXTENDED –0 != 1))
/*
 * Was the correct address bound?
 *
 * When using XTI, this test is unnecessary
 */

if (bind->addr.len != sizeof(int) ||
   *(int *)bind->addr.buf != SRV_ADDR) {
   fprintf(stderr, "t_bind bound wrong address\n");
   exit(4);
 }
 #endif
```

Like the client, the server first calls `t_open`(3NSL) to establish a transport endpoint with the desired transport provider. The endpoint, `listen_fd`, is used to listen for connect requests.

Next, the server binds its address to the endpoint. This address is used by each client to access the server. The second argument points to a `t_bind` structure that specifies the address to bind to the endpoint. The `t_bind` structure has the following format:

```
struct t_bind {
  struct netbuf addr;
  unsigned qlen;
}
```

Where `addr` describes the address to be bound, and `qlen` specifies the maximum number of outstanding connect requests. All XTI structure and constant definitions made visible for use by applications programs through `xti.h`. All TLI structure and constant definitions are in `tiuser.h`.

The address is specified in the `netbuf` structure with the following format:

```
struct netbuf {
  unsigned int maxlen;
  unsigned int len;
  char *buf;
}
```

Where `maxlen` specifies the maximum length of the buffer in bytes, `len` specifies the bytes of data in the buffer, and `buf` points to the buffer that contains the data.

In the `t_bind` structure, the data identifies a transport address. `qlen` specifies the maximum number of connect requests that can be queued. If the value of `qlen` is positive, the endpoint can be used to listen for connect requests. `t_bind`(3NSL) directs the transport provider to queue connect requests for the bound address immediately. The server must dequeue each connect request and accept or reject it. For a server that fully processes a single connect request and responds to it before receiving the next request, a value of `1` is appropriate for `qlen`. Servers that dequeue several connect requests before responding to any should specify a longer queue. The server in this example processes connect requests one at a time, so `qlen` is set to `1`.

`t_alloc`(3NSL) is called to allocate the `t_bind` structure. `t_alloc`(3NSL) has three arguments: a file descriptor of a transport endpoint; the identifier of the structure to allocate; and a flag that specifies which, if any, `netbuf` buffers to allocate. `T_ALL` specifies to allocate all `netbuf` buffers, and causes the `addr` buffer to be allocated in this example. Buffer size is determined automatically and stored in the `maxlen` field.

Each transport provider manages its address space differently. Some transport providers allow a single transport address to be bound to several transport endpoints, while others require a unique address per endpoint. XTI and TLI differ in some significant ways in providing the address binding.

In TLI, based on its rules, a provider determines if it can bind the requested address. If not, it chooses another valid address from its address space and binds it to the transport endpoint. The application program must check the bound address to ensure that it is the one previously advertised to clients. In XTI, if the provider determines it cannot bind to the requested address, it fails the `t_bind`(3NSL) request with an error.

If `t_bind`(3NSL) succeeds, the provider begins queueing connect requests, entering the next phase of communication.

Connection Establishment

XTI/TLI imposes different procedures in this phase for clients and servers. The client starts connection establishment by requesting a connection to a specified server using `t_connect`(3NSL). The server receives a client's request by calling `t_listen`(3NSL). The server must accept or reject the client's request. It calls `t_accept`(3NSL) to establish the connection, or `t_snddis`(3NSL) to reject the request. The client is notified of the result when `t_connect`(3NSL) returns.

TLI supports two facilities during connection establishment that might not be supported by all transport providers:

- Data transfer between the client and server when establishing the connection. The client can send data to the server when it requests a connection. This data is passed to the server by `t_listen`(3NSL). The server can send data to the client when it accepts or rejects the connection. The connect characteristic returned by `t_open`(3NSL) determines how much data, if any, two users can transfer during connect establishment.

- The negotiation of protocol options. The client can specify preferred protocol options to the transport provider and/or the remote user. XTI/TLI supports both local and remote option negotiation. Option negotiation is a protocol-specific capability.

These facilities produce protocol-dependent software (see "Guidelines to Protocol Independence" on page 112).

Client

The steps for the client to establish a connection are shown in Code Example 3–5.

CODE EXAMPLE 3–5 Client-to-Server Connection

```
if ((sndcall = (struct t_call *) t_alloc(fd, T_CALL, T_ADDR))
      == (struct t_call *) NULL) {
   t_error("t_alloc failed");
   exit(3);
}

/*
 * Because it assumes it knows the format of the provider's
 * address, this program is transport-dependent
 */
sndcall->addr.len = sizeof(int);
*(int *) sndcall->addr.buf = SRV_ADDR;
if (t_connect( fd, sndcall, (struct t_call *) NULL) == -1 ) {
   t_error("t_connect failed for fd");
   exit(4);
}
```

The `t_connect`(3NSL) call connects to the server. The first argument of `t_connect`(3NSL) identifies the client's endpoint, and the second argument points to a `t_call` structure that identifies the destination server. This structure has the following format:

```
struct t_call {
  struct netbuf addr;
  struct netbuf opt;
  struct netbuf udata;
  int sequence;
}
```

addr identifies the address of the server, opt specifies protocol-specific options to the connection, and udata identifies user data that can be sent with the connect request to the server. The sequence field has no meaning for t_connect(3NSL). In this example, only the server's address is passed.

t_alloc(3NSL) allocates the t_call structure dynamically. The third argument of t_alloc(3NSL) is T_ADDR, which specifies that the system needs to allocate a netbuf buffer. The server's address is then copied to buf, and len is set accordingly.

The third argument of t_connect(3NSL) can be used to return information about the newly established connection, and can return any user data sent by the server in its response to the connect request. The third argument here is set to NULL by the client. The connection is established on successful return of t_connect(3NSL). If the server rejects the connect request, t_connect(3NSL) sets t_errno to TLOOK.

Event Handling

The TLOOK error has special significance. TLOOK is set if an XTI/TLI routine is interrupted by an unexpected asynchronous transport event on the endpoint. TLOOK does not report an error with an XTI/TLI routine, but the normal processing of the routine is not done because of the pending event. The events defined by XTI/TLI are listed in Table 3–7.

TABLE 3–7 Asynchronous Endpoint Events

Name	Description
T_LISTEN	Connection request arrived at the transport endpoint
T_CONNECT	Confirmation of a previous connect request arrived (generated when a server accepts a connect request)
T_DATA	User data has arrived
T_EXDATA	Expedited user data arrived
T_DISCONNECT	Notice that an aborted connection or a rejected connect request arrived
T_ORDREL	A request for orderly release of a connection arrived
T_UDERR	Notice of an error in a previous datagram arrived. (See "Read/Write Interface" on page 91.)

The state table in “State Transitions” on page 104 shows which events can happen in each state. `t_look`(3NSL) lets a user determine what event has occurred if a `TLOOK` error is returned. In the example, if a connect request is rejected, the client exits.

Server

When the client calls `t_connect`(3NSL), a connect request is sent at the server's transport endpoint. For each client, the server accepts the connect request and spawns a process to service the connection.

```
if ((call = (struct t_call *) t_alloc(listen_fd, T_CALL, T_ALL))
      == (struct t_call *) NULL) {
   t_error("t_alloc of t_call structure failed");
   exit(5);
}
while(1) {
   if (t_listen( listen_fd, call) == –1) {
      t_error("t_listen failed for listen_fd");
      exit(6);
   }
   if ((conn_fd = accept_call(listen_fd, call)) != DISCONNECT)
      run_server(listen_fd);
}
```

The server allocates a `t_call` structure, then does a closed loop. The loop blocks on `t_listen`(3NSL) for a connect request. When a request arrives, the server calls `accept_call( )` to accept the connect request. `accept_call` accepts the connection on an alternate transport endpoint (as discussed below) and returns the handle of that endpoint. (*conn_fd* is a global variable.) Because the connection is accepted on an alternate endpoint, the server can continue to listen on the original endpoint. If the call is accepted without error, `run_server` spawns a process to service the connection.

XTI/TLI supports an asynchronous mode for these routines that prevents a process from blocking. See “Advanced Topics” on page 93.

When a connect request arrives, the server calls `accept_call` to accept the client's request, as Code Example 3–6 shows.

Note - It is implicitly assumed that this server only needs to handle a single connection request at a time. This is not normally true of a server. The code required to handle multiple simultaneous connection requests is complicated because of XTI/TLI event mechanisms. See “Advanced Programming Example” on page 94 for such a server.

CODE EXAMPLE 3–6 `accept_call` Function

```
accept_call(listen_fd, call)
int listen_fd;
struct t_call *call;
{
```

```
    int resfd;

    if ((resfd = t_open("/dev/exmp", O_RDWR, (struct t_info *) NULL))
          == -1) {
       t_error("t_open for responding fd failed");
       exit(7);
   }
    if (t_bind(resfd,(struct t_bind *) NULL, (struct t_bind *NULL))
          == -1) {
       t_error("t_bind for responding fd failed");
       exit(8);
    }
    if (t_accept(listen_fd, resfd, call) == -1) {
       if (t_errno == TLOOK) {        /* must be a disconnect */
          if (t_rcvdis(listen_fd,(struct t_discon *) NULL) == -1) {
             t_error("t_rcvdis failed for listen_fd");
             exit(9);
          }
          if (t_close(resfd) == -1) {
             t_error("t_close failed for responding fd");
             exit(10);
          }
          /* go back up and listen for other calls */
          return(DISCONNECT);
       }
       t_error("t_accept failed");
       exit(11);
    }
    return(resfd);
}
```

`accept_call()` has two arguments:

listen_fd	The file handle of the transport endpoint where the connect request arrived.
call	Points to a `t_call` structure that contains all information associated with the connect request

The server first opens another transport endpoint by opening the clone device special file of the transport provider and binding an address. A `NULL` specifies not to return the address bound by the provider. The new transport endpoint, *resfd*, accepts the client's connect request.

The first two arguments of `t_accept`(3NSL) specify the listening transport endpoint and the endpoint where the connection is accepted, respectively. Accepting a connection on the listening endpoint prevents other clients from accessing the server for the duration of the connection.

The third argument of `t_accept`(3NSL) points to the `t_call` structure containing the connect request. This structure should contain the address of the calling user and the sequence number returned by `t_listen`(3NSL). The sequence number is significant if the server queues multiple connect requests. The "Advanced Topics" on page 93 shows an example of this. The `t_call` structure also identifies protocol

options and user data to pass to the client. Because this transport provider does not support protocol options or the transfer of user data during connection, the `t_call` structure returned by `t_listen`(3NSL) is passed without change to `t_accept`(3NSL).

The example is simplified. The server exits if either the `t_open`(3NSL) or `t_bind`(3NSL) call fails. `exit`(2) closes the transport endpoint of `listen_fd`, causing a disconnect request to be sent to the client. The client's `t_connect`(3NSL) call fails, setting `t_errno` to `TLOOK`.

`t_accept`(3NSL) can fail if an asynchronous event occurs on the listening endpoint before the connection is accepted, and `t_errno` is set to `TLOOK`. Table 3–8 shows that only a disconnect request can be sent in this state with only one queued connect request. This event can happen if the client undoes a previous connect request. If a disconnect request arrives, the server must respond by calling `t_rcvdis`(3NSL). This routine argument is a pointer to a `t_discon` structure, which is used to retrieve the data of the disconnect request. In this example, the server passes a `NULL`.

After receiving a disconnect request, `accept_call` closes the responding transport endpoint and returns `DISCONNECT`, which informs the server that the connection was disconnected by the client. The server then listens for further connect requests.

Figure 3–4 illustrates how the server establishes connections:

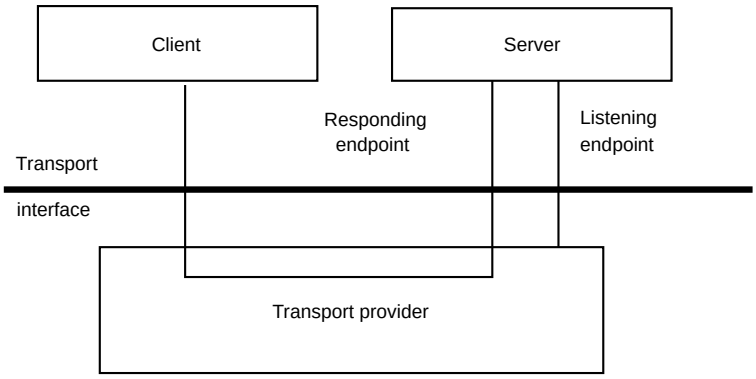

Figure 3–4 Listening and Responding Transport Endpoints

The transport connection is established on the new responding endpoint, and the listening endpoint is freed to retrieve further connect requests.

Data Transfer

After the connection is established, both the client and the server can transfer data through the connection using `t_snd`(3NSL) and `t_rcv`(3NSL). XTI/TLI does not

differentiate the client from the server from this point on. Either user can send data, receive data, or release the connection.

The two classes of data on a transport connection are:

1. Normal data
2. Expedited data

Expedited data is for urgent data. The exact semantics of expedited data vary between transport providers. Not all transport protocols support expedited data (see `t_open`(3NSL)).

Most connection-oriented mode protocols transfer data in byte streams. “Byte stream” implies no message boundaries in data sent over a connection. Some transport protocols preserve message boundaries over a transport connection. This service is supported by XTI/TLI, but protocol-independent software must not rely on it.

The message boundaries are invoked by the `T_MORE` flag of `t_snd`(3NSL) and `t_rcv`(3NSL). The messages, called transport service data units (TSDU), can be transferred between two transport users as distinct units. The maximum message size is defined by the underlying transport protocol. Get the message size through `t_open`(3NSL) or `t_getinfo`(3NSL).

You can send a message in multiple units. Set the `T_MORE` flag on every `t_snd`(3NSL) call, except the last to send a message in multiple units. The flag specifies that the data in the current and the next `t_snd`(3NSL) calls are a logical unit. Send the last message unit with `T_MORE` turned off to specify the end of the logical unit.

Similarly, a logical unit can be sent in multiple units. If `t_rcvv`(3NSL) returns with the `T_MORE` flag set, the user must call `t_rcvv`(3NSL) again to receive the rest of the message. The last unit in the message is identified by a call to `t_rcvv`(3NSL) that does not set `T_MORE`.

The `T_MORE` flag implies nothing about how the data is packaged below XTI/TLI or how the data is delivered to the remote user. Each transport protocol, and each implementation of a protocol, can package and deliver the data differently.

For example, if a user sends a complete message in a single call to `t_snd`(3NSL)`t_snd`, there is no guarantee that the transport provider delivers the data in a single unit to the receiving user. Similarly, a message transmitted in two units can be delivered in a single unit to the remote transport user.

If supported by the transport, the message boundaries are preserved only by setting the value of `T_MORE` for `t_snd`(3NSL) and testing it after `t_rcvv`(3NSL). This guarantees that the receiver sees a message with the same contents and message boundaries as was sent.

Client

The example server transfers a log file to the client over the transport connection. The client receives the data and writes it to its standard output file. A byte stream interface is used by the client and server, with no message boundaries. The client receives data by the following:

```
while ((nbytes = t_rcv(fd, buf, nbytes, &flags))!= –1){
   if (fwrite(buf, 1, nbytes, stdout) == –1) {
      fprintf(stderr, "fwrite failed\n");
      exit(5);
   }
}
```

The client repeatedly calls `t_rcvv`(3NSL) to receive incoming data. `t_rcvv`(3NSL) blocks until data arrives. `t_rcvv`(3NSL) writes up to *nbytes* of the data available into *buf* and returns the number of bytes buffered. The client writes the data to standard output and continues. The data transfer loop ends when `t_rcvv`(3NSL) fails. `t_rcvv`(3NSL) fails when an orderly release or disconnect request arrives. If `fwrite`(3C) fails for any reason, the client exits, which closes the transport endpoint. If the transport endpoint is closed (either by `exit`(2) or `t_close`(3NSL)) during data transfer, the connection is aborted and the remote user receives a disconnect request.

Server

The server manages its data transfer by spawning a child process to send the data to the client. The parent process continues the loop to listen for more connect requests. `run_server` is called by the server to spawn this child process, as shown in Code Example 3–7.

CODE EXAMPLE 3–7 Spawning Child Process to Loopback and Listen

```
connrelease()
{
   /* conn_fd is global because needed here */
   if (t_look(conn_fd) == T_DISCONNECT) {
      fprintf(stderr, ‘‘connection aborted\n’’);
      exit(12);
   }
   /* else orderly release request - normal exit */
   exit(0);
}
run_server(listen_fd)
int listen_fd;
{
   int nbytes;
   FILE *logfp;                    /* file pointer to log file */
   char buf[1024];

   switch(fork()) {
   case –1:
      perror("fork failed");
```

```
        exit(20);
    default:         /* parent */
        /* close conn_fd and then go up and listen again*/
        if (t_close(conn_fd) == -1) {
            t_error("t_close failed for conn_fd");
            exit(21);
        }
        return;
    case 0:                         /* child */
        /* close listen_fd and do service */
        if (t_close(listen_fd) == -1) {
            t_error("t_close failed for listen_fd");
            exit(22);
        }
        if ((logfp = fopen("logfile", "r")) == (FILE *) NULL) {
            perror("cannot open logfile");
            exit(23);
        }
        signal(SIGPOLL, connrelease);
        if (ioctl(conn_fd, I_SETSIG, S_INPUT) == -1) {
            perror("ioctl I_SETSIG failed");
            exit(24);
        }
        if (t_look(conn_fd) != 0){      /*disconnect there?*/
            fprintf(stderr, "t_look: unexpected event\n");
            exit(25);
        }
        while ((nbytes = fread(buf, 1, 1024, logfp)) > 0)
            if (t_snd(conn_fd, buf, nbytes, 0) == -1) {
                t_error("t_snd failed");
                exit(26);
            }
```

After the fork, the parent process returns to the main listening loop. The child process manages the newly established transport connection. If the fork fails, `exit`(2) closes both transport endpoints, sending a disconnect request to the client, and the client's `t_connect`(3NSL) call fails.

The server process reads 1024 bytes of the log file at a time and sends the data to the client using `t_snd`(3NSL). *buf* points to the start of the data buffer, and *nbytes* specifies the number of bytes to transmit. The fourth argument can be zero or one of the two optional flags below:

- `T_EXPEDITED` specifies that the data is expedited.
- `T_MORE` specifies that the next block continues the message in this block.

Neither flag is set by the server in this example.

If the user floods the transport provider with data, `t_snd`(3NSL) blocks until enough data is removed from the transport.

`t_snd`(3NSL) does not look for a disconnect request (showing that the connection was broken). If the connection is aborted, the server should be notified, since data can be lost. One solution is to call `t_look`(3NSL) to check for incoming events before each `t_snd`(3NSL) call or after a `t_snd`(3NSL) failure. The example has a cleaner solution. The `I_SETSIG ioctl`(2) lets a user request a signal when a

specified event occurs. See the `streamio`(7I) manpage. `S_INPUT` causes a signal to be sent to the user process when any input arrives at the endpoint `conn_fd`. If a disconnect request arrives, the signal-catching routine (`connrelease`) prints an error message and exits.

If the server alternates `t_snd`(3NSL) and `t_rcv`(3NSL) calls, it can use `t_rcv`(3NSL) to recognize an incoming disconnect request.

Connection Release

At any time during data transfer, either user can release the transport connection and end the conversation. There are two forms of connection release.

- The first way, abortive release, breaks the connection immediately and discards any data that has not been delivered to the destination user.

  Either user can call `t_snddis`(3NSL) to perform an abortive release. The transport provider can abort a connection if a problem occurs below XTI/TLI. `t_snddis`(3NSL) lets a user send data to the remote user when aborting a connection. The abortive release is supported by all transport providers, the ability to send data when aborting a connection is not.

  When the remote user is notified of the aborted connection, call `t_rcvdis`(3NSL) to receive the disconnect request. The call returns a code that identifies why the connection was aborted, and returns any data that can have accompanied the disconnect request (if the abort was initiated by the remote user). The reason code is specific to the underlying transport protocol, and should not be interpreted by protocol-independent software.

- The second way, orderly release, ends a connection so that no data is lost. All transport providers must support the abortive release procedure, but orderly release is an option not supported by all connection-oriented protocols.

See “Transport Selection” on page 120 for information on how to select a transport that supports orderly release.

Server

This example assumes that the transport provider supports orderly release. When all the data has been sent by the server, the connection is released as follows:

```
if (t_sndrel(conn_fd) == –1) {
   t_error(‘‘t_sndrel failed’’);
   exit(27);
}
pause(); /* until orderly release request arrives */
```

Orderly release requires two steps by each user. The server can call `t_sndrel`(3NSL). This routine sends a disconnect request. When the client receives

the request, it can continue sending data back to the server. When all data have been sent, the client calls t_sndrel(3NSL) to send a disconnect request back. The connection is released only after both users have received a disconnect request.

In this example, data is transferred only from the server to the client. So there is no provision to receive data from the client after the server initiates release. The server calls pause(2) after initiating the release.

The client responds with its orderly release request, which generates a signal caught by connrelease( ). (In Code Example 3–7, the server issued an I_SETSIG ioctl(2) to generate a signal on any incoming event.) The only XTI/TLI event possible in this state is a disconnect request or an orderly release request, so connrelease exits normally when the orderly release request arrives. exit(2) from connrelease closes the transport endpoint and frees the bound address. To close a transport endpoint without exiting, call t_close(3NSL).

Client

The client releases the connection similar to the way the server releases it. The client processes incoming data until t_rcv(3NSL) fails. When the server releases the connection (using either t_snddis(3NSL) or t_sndrel(3NSL)), t_rcv(3NSL) fails and sets t_errno to TLOOK. The client then processes the connection release as follows:

```
if ((t_errno == TLOOK) && (t_look(fd) == T_ORDREL)) {
   if (t_rcvrel(fd) == –1) {
      t_error(‘‘t_rcvrel failed’’);
      exit(6);
  }
  if (t_sndrel(fd) == –1) {
      t_error(‘‘t_sndrel failed’’);
      exit(7);
  }
  exit(0);
 }
```

Each event on the client’s transport endpoint is checked for an orderly release request. When one is received, the client calls t_rcvrel(3NSL) to process the request and t_sndrel(3NSL) to send the response release request. The client then exits, closing its transport endpoint.

If a transport provider does not support the orderly release, use abortive release with t_snddis(3NSL) and t_rcvdis(3NSL). Each user must take steps to prevent data loss. For example, use a special byte pattern in the data stream to indicate the end of a conversation.

Read / Write Interface

A user might want to establish a transport connection using `exec`(2) on an existing program (such as `/usr/bin/cat`) to process the data as it arrives over the connection. Existing programs use `read`(2) and `write`(2). XTI / TLI does not directly support a read / write interface to a transport provider, but one is available. The interface lets you issue `read`(2) and `write`(2) calls over a transport connection in the data transfer phase. This section describes the read / write interface to the connection mode service of XTI / TLI. This interface is not available with the connectionless mode service.

The read / write interface is presented using the client example (with modifications) of "Connection Mode Service" on page 75. The clients are identical until the data transfer phase. Then the client uses the read / write interface and `cat`(1) to process incoming data. `cat`(1) is run without change over the transport connection. Only the differences between this client and that of the client in Code Example 3–3are shown in Code Example 3–8.

CODE EXAMPLE 3–8 Read / Write Interface

```
#include <stropts.h>
   .
   ./*
     Same local management and connection establishment steps.
     */
   .
   if (ioctl(fd, I_PUSH, "tirdwr") == –1) {
      perror(‘‘I_PUSH of tirdwr failed’’);
      exit(5);
  }
   close(0);
   dup(fd);
   execl(‘‘/usr/bin/cat’’, ‘‘/usr/bin/cat’’, (char *) 0);
   perror(‘‘exec of /usr/bin/cat failed’’);
   exit(6);
}
```

The client invokes the read / write interface by pushing `tirdwr` onto the stream associated with the transport endpoint. See `I_PUSH` in `streamio`(7I). `tirdwr` converts XTI / TLI above the transport provider into a pure read / write interface. With the module in place, the client calls `close`(2) and `dup`(2) to establish the transport endpoint as its standard input file, and uses `/usr/bin/cat` to process the input.

By pushing `tirdwr` onto the transport provider, XTI / TLI is changed. The semantics of `read`(2) and `write`(2) must be used, and message boundaries are not preserved. `tirdwr` can be popped from the transport provider to restore XTI / TLI semantics (see `I_POP` in `streamio`(7I).

Caution - The `tirdwr` module can only be pushed onto a stream when the transport endpoint is in the data transfer phase. After the module is pushed, the user cannot call any XTI/TLI routines. If an XTI/TLI routine is invoked, `tirdwr` generates a fatal protocol error, `EPROTO`, on the stream, rendering it unusable. If you then pop the `tirdwr` module off the stream, the transport connection is aborted. See `I_POP` in `streamio`(7I).

Write

Send data over the transport connection with `write`(2). `tirdwr` passes data through to the transport provider. If you send a zero-length data packet, which the mechanism allows, `tirdwr` discards the message. If the transport connection is aborted—for example, because the remote user aborts the connection using `t_snddis`(3NSL)—a hang-up condition is generated on the stream, further `write`(2) calls fail, and `errno` is set to `ENXIO`. You can still retrieve any available data after a hang-up.

Read

Receive data that arrives at the transport connection with `read`(2). `tirdwr`, which passes data from the transport provider. Any other event or request passed to the user from the provider is processed by `tirdwr` as follows:

- `read`(2) cannot identify expedited data to the user. If an expedited data request is received, `tirdwr` generates a fatal protocol error, `EPROTO`, on the stream. The error causes further system calls to fail. Do not use `read`(2) to receive expedited data.
- `tirdwr` discards an abortive disconnect request and generates a hang-up condition on the stream. Subsequent `read`(2) calls retrieve any remaining data, then return zero for all further calls (indicating end of file).
- `tirdwr` discards an orderly release request and delivers a zero-length message to the user. As described in `read`(2), this notifies the user of end of file by returning 0.
- If any other XTI/TLI request is received, `tirdwr` generates a fatal protocol error, `EPROTO`, on the stream. This causes further system calls to fail. If a user pushes `tirdwr` onto a stream after the connection has been established, no request is generated.

Close

With `tirdwr` on a stream, you can send and receive data over a transport connection for the duration of the connection. Either user can terminate the connection by closing the file descriptor associated with the transport endpoint or by popping the `tirdwr` module off the stream. In either case, `tirdwr` does the following:

- If an orderly release request was previously received by `tirdwr`, it is passed to the transport provider to complete the orderly release of the connection. The remote user who initiated the orderly release procedure receives the expected request when data transfer completes.
- If a disconnect request was previously received by `tirdwr`, no special action is taken.
- If neither an orderly release nor a disconnect request was previously received by `tirdwr`, a disconnect request is passed to the transport provider to abort the connection.
- If an error previously occurred on the stream and a disconnect request has not been received by `tirdwr`, a disconnect request is passed to the transport provider.

A process cannot initiate an orderly release after `tirdwr` is pushed onto a stream. `tirdwr` handles an orderly release if it is initiated by the user on the other side of a transport connection. If the client in this section is communicating with the server program in “Connection Mode Service” on page 75, the server terminates the transfer of data with an orderly release request. The server then waits for the corresponding request from the client. At that point, the client exits and the transport endpoint is closed. When the file descriptor is closed, `tirdwr` initiates the orderly release request from the client’s side of the connection. This generates the request that the server is blocked on.

Some protocols, like TCP, require this orderly release to ensure that the data is delivered intact.

Advanced Topics

This section presents additional XTI/TLI concepts:

- An optional nonblocking (asynchronous) mode for some library calls
- How to set and get TCP and UDP options under XTI/TLI
- A program example of a server supporting multiple outstanding connect requests and operating in an event-driven manner

Asynchronous Execution Mode

Many XTI/TLI library routines block to wait for an incoming event. However, some time-critical applications should not block for any reason. An application can do local processing while waiting for some asynchronous XTI/TLI event.

Asynchronous processing of XTI/TLI events is available to applications through the combination of asynchronous features and the non-blocking mode of XTI/TLI library routines. Use of the `poll`(2) system call and the `I_SETSIG ioctl`(2) command to process events asynchronously is described in *ONC+ Developer's Guide*.

Each XTI/TLI routine that blocks for an event can be run in a special non-blocking mode. For example, `t_listen`(3NSL) normally blocks for a connect request. A server can periodically poll a transport endpoint for queued connect requests by calling `t_listen`(3NSL) in the non-blocking (or asynchronous) mode. The asynchronous mode is enabled by setting `O_NDELAY` or `O_NONBLOCK` in the file descriptor. These modes can be set as a flag through `t_open`(3NSL), or by calling `fcntl`(2) before calling the XTI/TLI routine. `fcntl`(2) enables or disables this mode at any time. All program examples in this chapter use the default synchronous processing mode.

`O_NDELAY` or `O_NONBLOCK` affect each XTI/TLI routine differently. You will need to determine the exact semantics of `O_NDELAY` or `O_NONBLOCK` for a particular routine.

Advanced Programming Example

The following example demonstrates two important concepts. The first is a server's ability to manage multiple outstanding connect requests. The second is event-driven use of XTI/TLI and the system call interface.

The server example in Code Example 3–4 supports only one outstanding connect request, but XTI/TLI lets a server manage multiple outstanding connect requests. One reason to receive several simultaneous connect requests is to prioritize the clients. A server can receive several connect requests, and accept them in an order based on the priority of each client.

The second reason for handling several outstanding connect requests is the limits of single-threaded processing. Depending on the transport provider, while a server processes one connect request, other clients find it busy. If multiple connect requests are processed simultaneously, the server will be found busy only if more than the maximum number of clients try to call the server simultaneously.

The server example is event-driven: the process polls a transport endpoint for incoming XTI/TLI events, and takes the appropriate actions for the event received. The example demonstrates the ability to poll multiple transport endpoints for incoming events.

The definitions and endpoint establishment functions of Code Example 3–9 are similar to those of the server example in Code Example 3–4.

CODE EXAMPLE 3–9 Endpoint Establishment (Convertible to Multiple Connections)

```
#include <tiuser.h>
#include <fcntl.h>
#include <stdio.h>
#include <poll.h>
#include <stropts.h>
#include <signal.h>

#define NUM_FDS 1
#define MAX_CONN_IND 4
#define SRV_ADDR 1                /* server's well known address */

int conn_fd;                      /* server connection here */
extern int t_errno;
/* holds connect requests */
struct t_call *calls[NUM_FDS][MAX_CONN_IND];

main()
{
   struct pollfd pollfds[NUM_FDS];
   struct t_bind *bind;
   int i;

   /*
    * Only opening and binding one transport endpoint, but more can
    * be supported
    */
   if ((pollfds[0].fd = t_open(''/dev/tivc'', O_RDWR,
         (struct t_info *) NULL)) == -1) {
      t_error(''t_open failed'');
      exit(1);
   }
   if ((bind = (struct t_bind *) t_alloc(pollfds[0].fd, T_BIND,
         T_ALL)) == (struct t_bind *) NULL) {
      t_error(''t_alloc of t_bind structure failed'');
      exit(2);
   }
   bind->qlen = MAX_CONN_IND;
   bind->addr.len = sizeof(int);
   *(int *) bind->addr.buf = SRV_ADDR;
   if (t_bind(pollfds[0].fd, bind, bind) == -1) {
      t_error(''t_bind failed'');
      exit(3);
   }
   /* Was the correct address bound? */
   if (bind->addr.len != sizeof(int) ||
      *(int *)bind->addr.buf != SRV_ADDR) {
      fprintf(stderr, ''t_bind bound wrong address\n'');
      exit(4);
   }
}
```

The file descriptor returned by `t_open`(3NSL) is stored in a `pollfd` structure that controls polling the transport endpoints for incoming data. See `poll`(2). Only one transport endpoint is established in this example. However, the remainder of the example is written to manage multiple transport endpoints. Several endpoints could be supported with minor changes to Code Example 3–9.

This server sets `qlen` to a value greater than `1` for `t_bind`(3NSL). This specifies that the server queues multiple outstanding connect requests. The server accepts the current connect request before accepting additional connect requests. This example can queue up to `MAX_CONN_IND` connect requests. The transport provider can negotiate the value of `qlen` smaller if it cannot support `MAX_CONN_IND` outstanding connect requests.

After the server has bound its address and is ready to process connect requests, it behaves as shown in Code Example 3–10.

CODE EXAMPLE 3–10 Processing Connection Requests

```
pollfds[0].events = POLLIN;

while (TRUE) {
 if (poll(pollfds, NUM_FDS, –1) == –1) {
   perror(‘‘poll failed’’);
   exit(5);
 }
 for (i = 0; i < NUM_FDS; i++) {
   switch (pollfds[i].revents) {
      default:
         perror(‘‘poll returned error event’’);
      exit(6);
      case 0:
         continue;
      case POLLIN:
         do_event(i, pollfds[i].fd);
         service_conn_ind(i, pollfds[i].fd);
    }
   }
}
```

The `events` field of the `pollfd` structure is set to `POLLIN`, which notifies the server of any incoming XTI/TLI events. The server then enters an infinite loop in which it polls the transport endpoint(s) for events, and processes events as they occur.

The `poll`(2) call blocks indefinitely for an incoming event. On return, each entry (one per transport endpoint) is checked for a new event. If `revents` is `0`, no event has occurred on the endpoint and the server continues to the next endpoint. If `revents` is `POLLIN`, there is an event on the endpoint. `do_event` is called to process the event. Any other value in `revents` indicates an error on the endpoint, and the server exits. With multiple endpoints, it is better for the server to close this descriptor and continue.

For each iteration of the loop, `service_conn_ind` is called to process any outstanding connect requests. If another connect request is pending, `service_conn_ind` saves the new connect request and responds to it later.

The `do_event` in Code Example 3–11 is called to process an incoming event.

CODE EXAMPLE 3–11 Event Processing Routine

```
do_event( slot, fd)
int slot;
int fd;
{
   struct t_discon *discon;
   int i;

   switch (t_look(fd)) {
   default:
      fprintf(stderr, "t_look: unexpected event\n");
      exit(7);
   case T_ERROR:
      fprintf(stderr, "t_look returned T_ERROR event\n");
      exit(8);
   case –1:
      t_error("t_look failed");
      exit(9);
   case 0:
      /* since POLLIN returned, this should not happen */
      fprintf(stderr,"t_look returned no event\n");
      exit(10);
   case T_LISTEN:
      /* find free element in calls array */
      for (i = 0; i < MAX_CONN_IND; i++) {
         if (calls[slot][i] == (struct t_call *) NULL)
            break;
      }
      if ((calls[slot][i] = (struct t_call *) t_alloc( fd, T_CALL,
               T_ALL)) == (struct t_call *) NULL) {
         t_error("t_alloc of t_call structure failed");
         exit(11);
      }
      if (t_listen(fd, calls[slot][i] ) == –1) {
         t_error("t_listen failed");
         exit(12);
      }
      break;
   case T_DISCONNECT:
      discon = (struct t_discon *) t_alloc(fd, T_DIS, T_ALL);
      if (discon == (struct t_discon *) NULL) {
         t_error("t_alloc of t_discon structure failed");
         exit(13)
      }
      if(t_rcvdis( fd, discon) == –1) {
         t_error("t_rcvdis failed");
         exit(14);
      }
      /* find call ind in array and delete it */
      for (i = 0; i < MAX_CONN_IND; i++) {
         if (discon->sequence == calls[slot][i]->sequence) {
            t_free(calls[slot][i], T_CALL);
            calls[slot][i] = (struct t_call *) NULL;
         }
      }
      t_free(discon, T_DIS);
      break;
   }
}
```

The arguments are a number (*slot*) and a file descriptor (*fd*). *slot* is the index into the global array `calls` which has an entry for each transport endpoint. Each entry is an array of `t_call` structures that hold incoming connect requests for the endpoint.

`do_event` calls `t_look`(3NSL) to identify the XTI/TLI event on the endpoint specified by *fd*. If the event is a connect request (`T_LISTEN` event) or disconnect request (`T_DISCONNECT` event), the event is processed. Otherwise, the server prints an error message and exits.

For connect requests, `do_event` scans the array of outstanding connect requests for the first free entry. A `t_call` structure is allocated for the entry, and the connect request is received by `t_listen`(3NSL). The array is large enough to hold the maximum number of outstanding connect requests. The processing of the connect request is deferred.

A disconnect request must correspond to an earlier connect request. `do_event` allocates a `t_discon` structure to receive the request. This structure has the following fields:

```
struct t_discon {
  struct netbuf udata;
  int reason;
  int sequence;
}
```

`udata` contains any user data sent with the disconnect request. `reason` contains a protocol-specific disconnect reason code. `sequence` identifies the connect request that matches the disconnect request.

`t_rcvdis`(3NSL) is called to receive the disconnect request. The array of connect requests is scanned for one that contains the sequence number that matches the `sequence` number in the disconnect request. When the connect request is found, its structure is freed and the entry is set to `NULL`.

When an event is found on a transport endpoint, `service_conn_ind` is called to process all queued connect requests on the endpoint, as Code Example 3–12 shows.

CODE EXAMPLE 3–12 Process All Connect Requests

```
service_conn_ind(slot, fd)
{
   int i;

 for (i = 0; i < MAX_CONN_IND; i++) {
      if (calls[slot][i] == (struct t_call *) NULL)
         continue;
      if((conn_fd = t_open( ‘‘/dev/tivc’’, O_RDWR,
            (struct t_info *) NULL)) == –1) {
         t_error("open failed");
         exit(15);
      }
      if (t_bind(conn_fd, (struct t_bind *) NULL,
            (struct t_bind *) NULL) == –1) {
         t_error("t_bind failed");
         exit(16);
```

```
      }
      if (t_accept(fd, conn_fd, calls[slot][i]) == -1) {
         if (t_errno == TLOOK) {
            t_close(conn_fd);
            return;
         }
         t_error("t_accept failed");
         exit(167);
      }
      t_free(calls[slot][i], T_CALL);
      calls[slot][i] = (struct t_call *) NULL;
      run_server(fd);
   }
}
```

For each transport endpoint, the array of outstanding connect requests is scanned. For each request, the server opens a responding transport endpoint, binds an address to the endpoint, and accepts the connection on the endpoint. If another event (connect request or disconnect request) arrives before the current request is accepted, `t_accept`(3NSL) fails and sets `t_errno` to `TLOOK`. (You cannot accept an outstanding connect request if any pending connect request events or disconnect request events exist on the transport endpoint.)

If this error occurs, the responding transport endpoint is closed and `service_conn_ind` returns immediately (saving the current connect request for later processing). This causes the server's main processing loop to be entered, and the new event is discovered by the next call to `poll`(2). In this way, multiple connect requests can be queued by the user.

Eventually, all events are processed, and `service_conn_ind` is able to accept each connect request in turn. After the connection has been established, the `run_server` routine used by the server in the Code Example 3–5 is called to manage the data transfer.

Asynchronous Networking

This section discusses the techniques of asynchronous network communication using XTI/TLI for real-time applications. SunOS provides support for asynchronous network processing of XTI/TLI events using a combination of STREAMS asynchronous features and the non-blocking mode of the XTI/TLI library routines.

Networking Programming Models

Like file and device I/O, network transfers can be done synchronously or asynchronously with process service requests.

Synchronous Networking

Synchronous networking proceeds similar to synchronous file and device I/O. Like the `write`(2) function, the request to send returns after buffering the message, but might suspend the calling process if buffer space is not immediately available. Like the `read`(2) function, a request to receive suspends execution of the calling process until data arrives to satisfy the request. Because SunOS provides no guaranteed bounds for transport services, synchronous networking is inappropriate for processes that must have real-time behavior with respect to other devices.

Asynchronous Networking

Asynchronous networking is provided by non-blocking service requests. Additionally, applications can request asynchronous notification when a connection might be established, when data might be sent, or when data might be received.

Asynchronous Connectionless-Mode Service

Asynchronous connectionless mode networking is conducted by configuring the endpoint for non-blocking service, and either polling for or receiving asynchronous notification when data might be transferred. If asynchronous notification is used, the actual receipt of data typically takes place within a signal handler.

Making the Endpoint Asynchronous

After the endpoint has been established using `t_open`(3NSL), and its identity established using `t_bind`(3NSL), the endpoint can be configured for asynchronous service. This is done by using the `fcntl`(2) function to set the `O_NONBLOCK` flag on the endpoint. Thereafter, calls to `t_sndudata`(3NSL) for which no buffer space is immediately available return `-1` with `t_errno` set to `TFLOW`. Likewise, calls to `t_rcvudata`(3NSL) for which no data are available return `-1` with `t_errno` set to `TNODATA`.

Asynchronous Network Transfers

Although an application can use the `poll`(2) function to check periodically for the arrival of data or to wait for the receipt of data on an endpoint, it might be necessary to receive asynchronous notification when data has arrived. This can be done by using the `ioctl`(2) function with the `I_SETSIG` command to request that a `SIGPOLL` signal be sent to the process upon receipt of data at the endpoint. Applications should check for the possibility of multiple messages causing a single signal.

In the following example, `protocol` is the name of the application-chosen transport protocol.

```
#include <sys/types.h>
#include <tiuser.h>
#include <signal.h>
#include <stropts.h>

int    fd;
struct t_bind    *bind;
void    sigpoll(int);

 fd = t_open(protocol, O_RDWR, (struct t_info *) NULL);

 bind = (struct t_bind *) t_alloc(fd, T_BIND, T_ADDR);
 ...     /* set up binding address */
 t_bind(fd, bind, bin

 /* make endpoint non-blocking */
 fcntl(fd, F_SETFL, fcntl(fd, F_GETFL) | O_NONBLOCK);

 /* establish signal handler for SIGPOLL */
 signal(SIGPOLL, sigpoll);

 /* request SIGPOLL signal when receive data is available */
 ioctl(fd, I_SETSIG, S_INPUT | S_HIPRI);

 ...

void sigpoll(int sig)
{
 int     flags;
 struct t_unitdata     ud;

 for (;;) {
  ... /* initialize ud */
  if (t_rcvudata(fd, &ud, &flags) < 0) {
   if (t_errno == TNODATA)
    break;  /* no more messages */
   ... /* process other error conditions */
 }
 ... /* process message in ud */
}
```

Asynchronous Connection-Mode Service

For connection-mode service, an application can arrange for not only the data transfer, but for the establishment of the connection itself to be done asynchronously. The sequence of operations depends on whether the process is attempting to connect to another process or is awaiting connection attempts.

Asynchronously Establishing a Connection

A process can attempt a connection and asynchronously complete the connection. The process first creates the connecting endpoint, and, using `fcntl`(2), configures

the endpoint for non-blocking operation. As with connectionless data transfers, the endpoint can also be configured for asynchronous notification upon completion of the connection and subsequent data transfers. The connecting process then uses the `t_connect`(3NSL) function to initiate setting up the transfer. Then the `t_rcvconnect`(3NSL) function is used to confirm the establishment of the connection.

Asynchronous Use of a Connection

To asynchronously await connections, a process first establishes a non-blocking endpoint bound to a service address. When either the result of `poll`(2) or an asynchronous notification indicates that a connection request has arrived, the process can get the connection request by using the `t_listen`(3NSL) function. To accept the connection, the process uses the `t_accept`(3NSL) function. The responding endpoint must be separately configured for asynchronous data transfers.

The following example illustrates how to request a connection asynchronously.

```
#include <tiuser.h>
int             fd;
struct t_call   *call;

 fd = .../* establish a non-blocking endpoint */

 call = (struct t_call *) t_alloc(fd, T_CALL, T_ADDR);
 .../* initialize call structure */
 t_connect(fd, call, call);

 /* connection request is now proceeding asynchronously */

 .../* receive indication that connection has been accepted */
 t_rcvconnect(fd, &call);
```

The following example illustrates listening for connections asynchronously.

```
#include <tiuser.h>
int             fd, res_fd;
struct t_call   call;

 fd = ... /* establish non-blocking endpoint */

 .../*receive indication that connection request has arrived
*/
 call = (struct t_call *) t_alloc(fd, T_CALL, T_ALL);
 t_listen(fd, &call);

 .../* determine whether or not to accept connection */
 res_fd = ... /* establish non-blocking endpoint for response
*/
 t_accept(fd, res_fd, call);
```

Asynchronous Open

Occasionally, an application might be required to dynamically open a regular file in a file system mounted from a remote host, or on a device whose initialization might be prolonged. However, while such an open is in progress, the application is unable to achieve real-time response to other events. Fortunately, SunOS provides a means of solving this problem by having a second process perform the actual open and then pass the file descriptor to the real-time process.

Transferring a File Descriptor

The STREAMS interface under SunOS provides a mechanism for passing an open file descriptor from one process to another. The process with the open file descriptor uses the `ioctl`(2) function with a command argument of `I_SENDFD`. The second process obtains the file descriptor by calling `ioctl`(2) with a command argument of `I_RECVFD`.

In this example, the parent process prints out information about the test file, and creates a pipe. Next, the parent creates a child process, which opens the test file, and passes the open file descriptor back to the parent through the pipe. The parent process then displays the status information on the new file descriptor.

CODE EXAMPLE 3–13 File Descriptor Transfer

```
#include <sys/types.h>
#include <sys/stat.h>
#include <fcntl.h>
#include <stropts.h>
#include <stdio.h>

#define TESTFILE "/dev/null"
main(int argc, char *argv[])
{
 int fd;
 int pipefd[2];
 struct stat statbuf;

 stat(TESTFILE, &statbuf);
 statout(TESTFILE, &statbuf);
 pipe(pipefd);
 if (fork() == 0) {
  close(pipefd[0]);
  sendfd(pipefd[1]);
 } else {
  close(pipefd[1])
  recvfd(pipefd[0]);
 }
}

sendfd(int p)
{
 int tfd;

 tfd = open(TESTFILE, O_RDWR);
```

```
 ioctl(p, I_SENDFD, tfd);
}

recvfd(int p)
{
 struct strrecvfd rfdbuf;
 struct stat statbuf;
 char   fdbuf[32];

 ioctl(p, I_RECVFD, &rfdbuf);
 fstat(rfdbuf.fd, &statbuf);
 sprintf(fdbuf, "recvfd=%d", rfdbuf.fd);
 statout(fdbuf, &statbuf);
}

statout(char *f, struct stat *s)
{
 printf("stat: from=%s mode=0%o, ino=%ld, dev=%lx, rdev=%lx\n",
  f, s->st_mode, s->st_ino, s->st_dev, s->st_rdev);
 fflush(stdout);
}
```

State Transitions

These tables describe all state transitions associated with XTI/TLI. First, however, the states and events are described.

XTI/TLI States

Table 3–8 defines the states used in XTI/TLI state transitions, along with the service types.

TABLE 3–8 XTI/TLI State Transitions and Service Types

State	Description	Service Type
T_UNINIT	Uninitialized – initial and final state of interface	T_COTS, T_COTS_ORD, T_CLTS
T_UNBND	Initialized but not bound	T_COTS, T_COTS_ORD, T_CLTS
T_IDLE	No connection established	T_COTS, T_COTS_ORD, T_CLTS

TABLE 3–8 XTI/TLI State Transitions and Service Types *(continued)*

State	Description	Service Type
T_OUTCON	Outgoing connection pending for client	T_COTS, T_COTS_ORD
T_INCON	Incoming connection pending for server	T_COTS, T_COTS_ORD
T_DATAXFER	Data transfer	T_COTS, T_COTS_ORD
T_OUTREL	Outgoing orderly release (waiting for orderly release request	T_COTS_ORD
T_INREL	Incoming orderly release (waiting to send orderly release request)	T_COTS_ORD

Outgoing Events

The outgoing events described in Table 3–9 correspond to the status returned from the specified transport routines, where these routines send a request or response to the transport provider. In the table, some events, such as 'accept', are distinguished by the context in which they occur. The context is based on the values of the following variables:

- *ocnt* – Count of outstanding connect requests
- *fd* – File descriptor of the current transport endpoint
- *resfd* – File descriptor of the transport endpoint where a connection is accepted

TABLE 3–9 Outgoing Events

Event	Description	Service Type
opened	Successful return of t_open(3NSL)	T_COTS, T_COTS_ORD, T_CLTS
bind	Successful return of t_bind(3NSL)	T_COTS, T_COTS_ORD, T_CLTS
optmgmt	Successful return of t_optmgmt(3NSL)	T_COTS, T_COTS_ORD, T_CLTS

TABLE 3–9 Outgoing Events *(continued)*

Event	Description	Service Type
unbind	Successful return of t_unbind(3NSL)	T_COTS, T_COTS_ORD, T_CLTS
closed	Successful return of t_close(3NSL)	T_COTS, T_COTS_ORD, T_CLT
connect1	Successful return of t_connect(3NSL) in synchronous mode	T_COTS, T_COTS_ORD
connect2	TNODATA error on t_connect(3NSL) in asynchronous mode, or TLOOK error due to a disconnect request arriving on the transport endpoint	T_COTS, T_COTS_ORD
accept1	Successful return of t_accept(3NSL) with ocnt == 1, fd == resfd	T_COTS, T_COTS_ORD
accept2	Successful return of t_accept(3NSL) with ocnt== 1, fd!= resfd	T_COTS, T_COTS_ORD
accept3	Successful return of t_accept(3NSL) with ocnt > 1	T_COTS, T_COTS_ORD
snd	Successful return of t_snd(3NSL)	T_COTS, T_COTS_ORD
snddis1	Successful return of t_snddis(3NSL) with ocnt <= 1	T_COTS, T_COTS_ORD
snddis2	Successful return of t_snddis(3NSL) with ocnt > 1	T_COTS, T_COTS_ORD
sndrel	Successful return of t_sndrel(3NSL)	T_COTS_ORD
sndudata	Successful return of t_sndudata(3NSL)	T_CLTS

Incoming Events

The incoming events correspond to the successful return of the specified routines. These routines return data or event information from the transport provider. The only incoming event not associated directly with the return of a routine is `pass_conn`, which occurs when a connection is transferred to another endpoint. The event occurs on the endpoint that is being passed the connection, although no XTI/TLI routine is called on the endpoint.

In Table 3–10, the `rcvdis` events are distinguished by the value of `ocnt`, the count of outstanding connect requests on the endpoint.

TABLE 3–10 Incoming Events

Event	Description	Service Type
`listen`	Successful return of `t_listen`(3NSL)	`T_COTS`, `T_COTS_ORD`
`rcvconnect`	Successful return of `t_rcvconnect`(3NSL)	`T_COTS`, `T_COTS_ORD`
`rcv`	Successful return of `t_rcv`(3NSL)	`T_COTS`, `T_COTS_ORD`
`rcvdis1`	Successful return of `t_rcvdis`(3NSL)`rcvdis1t_rcvdis( )`, `onct <= 0`	`T_COTS`, `T_COTS_ORD`
`rcvdis2`	Successful return of `t_rcvdis`(3NSL), `ocnt == 1`	`T_COTS`, `T_COTS_ORD`
`rcvdis3`	Successful return of `t_rcvdis`(3NSL) with `ocnt > 1`	`T_COTS`, `T_COTS_ORD`
`rcvrel`	Successful return of `t_rcvrel`(3NSL)	`T_COTS_ORD`
`rcvudata`	Successful return of `t_rcvudata`(3NSL)	`T_CLTS`
`rcvuderr`	Successful return of `t_rcvuderr`(3NSL)	`T_CLTS`
`pass_conn`	Receive a passed connection	`T_COTS`, `T_COTS_ORD`

TABLE 3–10 Incoming Events *(continued)*

Transport User Actions

Some state transitions (below) have a list of actions the transport user must take. Each action is represented by a digit from the list below:

- Set the count of outstanding connect requests to zero.
- Increment the count of outstanding connect requests.
- Decrement the count of outstanding connect requests.
- Pass a connection to another transport endpoint, as indicated in `t_accept`(3NSL).

State Tables

The tables describe the XTI/TLI state transitions. Each box contains the next state, given the current state (column) and the current event (row). An empty box is an invalid state/event combination. Each box can also have an action list. Actions must be done in the order specified in the box.

The following should be understood when studying the state tables:

- `t_close`(3NSL) causes an established connection to be terminated for a connection-oriented transport provider. The connection termination will be orderly or abortive, depending on the service type supported by the transport provider. See `t_getinfo`(3NSL).
- If a transport user issues a function out of sequence, the function fails and `t_errno` is set to `TOUTSTATE`. The state does not change.
- The error codes `TLOOK` or `TNODATA` after `t_connect`(3NSL) can result in state changes described in "Event Handling" on page 82. The state tables assume correct use of XTI/TLI.
- Any other transport error does not change the state, unless the manual page for the function says otherwise.
- The support functions `t_getinfo`(3NSL), `t_getstate`(3NSL), `t_alloc`(3NSL), `t_free`(3NSL), `t_sync`(3NSL), `t_look`(3NSL), and `t_error`(3NSL) are excluded from the state tables because they do not affect the state.

Table 3–11, Table 3–12, Table 3–13, and Table 3–14 show endpoint establishment, data transfer in connectionless mode, and connection establishment/connection release/ data transfer in connection mode.

TABLE 3–11 Connection Establishment State

Event/State	T_UNINIT	T_UNBND	T_IDLE
`opened`	T_UNBND		
`bind`		T_IDLE[1]	
`optmgmt` (TLI only)			T_IDLE
`unbind`			T_UNBND
`closed`		T_UNINIT	

TABLE 3–12 Connection Mode State—Part 1

Event/State	T_IDLE	T_OUTCON	T_INCON	T_DATAXFER
`connect1`	T_DATAXFER			
`connect2`	T_OUTCON			
`rcvconnect`		T_DATAXFER		
`listen`	T_INCON [2]		T_INCON [2]	
`accept1`			T_DATAXFER [3]	
`accept2`			T_IDLE [3] [4]	
`accept3`			T_INCON [3] [4]	
`snd`				T_DATAXFER
`rcv`				T_DATAXFER
`snddis1`		T_IDLE	T_IDLE [3]	T_IDLE

TABLE 3–12 Connection Mode State—Part 1 *(continued)*

Event/State	T_IDLE	T_OUTCON	T_INCON	T_DATAXFER
snddis2			T_INCON [3]	
rcvdis1		T_IDLE		T_IDLE
rcvdis2			T_IDLE [3]	
rcvdis3			T_INCON [3]	
sndrel				T_OUTREL
rcvrel				T_INREL
pass_conn	T_DATAXFER			
optmgmt	T_IDLE	T_OUTCON	T_INCON	T_DATAXFER
closed	T_UNINIT	T_UNINIT	T_UNINIT	T_UNINIT

TABLE 3–13 Connection Mode State—Part 2

Event/State	T_OUTREL	T_INREL	T_UNBND
connect1			
connect2			
rcvconnect			
listen			
accept1			
accept2			
accept3			

TABLE 3–13 Connection Mode State—Part 2 *(continued)*

Event/State	T_OUTREL	T_INREL	T_UNBND
snd		T_INREL	
rcv	T_OUTREL		
snddis1	T_IDLE	T_IDLE	
snddis2			
rcvdis1	T_IDLE	T_IDLE	
rcvdis2			
rcvdis3			
sndrel		T_IDLE	
rcvrel	T_IDLE		
pass_conn			T_DATAXFER
optmgmt	T_OUTREL	T_INREL	T_UNBND
closed	T_UNINIT	T_UNINIT	

TABLE 3–14 Connectionless Mode State

Event/State	T_IDLE
snudata	T_IDLE
rcvdata	T_IDLE
rcvuderr	T_IDLE

Guidelines to Protocol Independence

XTI/TLI's set of services, common to many transport protocols, offers protocol independence to applications. Not all transport protocols support all XTI/TLI services. If software must run in a variety of protocol environments, use only the common services. The following is a list of services that might not be common to all transport protocols.

1. In connection mode service, a transport service data unit (TSDU) might not be supported by all transport providers. Make no assumptions about preserving logical data boundaries across a connection.
2. Protocol and implementation specific service limits are returned by the `t_open`(3NSL) and `t_getinfo`(3NSL) routines. Use these limits to allocate buffers to store protocol-specific transport addresses and options.
3. Do not send user data with connect requests or disconnect requests, such as `t_connect`(3NSL) and `t_snddis`(3NSL). Not all transport protocols work this way.
4. The buffers in the `t_call` structure used for `t_listen`(3NSL) must be large enough to hold any data sent by the client during connection establishment. Use the `T_ALL` argument to `t_alloc`(3NSL) to set maximum buffer sizes to store the address, options, and user data for the current transport provider.
5. Do not specify a protocol address on `t_bind`(3NSL) on a client side endpoint. Let the transport provider assign an appropriate address to the transport endpoint. A server should retrieve its address for `t_bind`(3NSL) in such a way that it does not require knowledge of the transport provider's name space.
6. Do not make assumptions about formats of transport addresses. Transport addresses should not be constants in a program. Chapter 4 contains detailed information.
7. The reason codes associated with `t_rcvdis`(3NSL) are protocol-dependent. Do not interpret this information if protocol independence is important.
8. The `t_rcvuderr`(3NSL) error codes are protocol dependent. Do not interpret this information if protocol independence is a concern.
9. Do not code the names of devices into programs. The device node identifies a particular transport provider and is not protocol independent. See Chapter 4 for details.
10. Do not use the optional orderly release facility of the connection mode service—provided by `t_sndrel`(3NSL) and `t_rcvrel`(3NSL)—in programs targeted for multiple protocol environments. This facility is not supported by all connection-based transport protocols. Its use can prevent programs from successfully communicating with open systems.

XTI/TLI Versus Socket Interfaces

XTI/TLI and sockets are different methods of handling the same tasks. Mostly, they provide mechanisms and services that are functionally similar. They do not provide one-to-one compatibility of routines or low-level services. Observe the similarities and differences between the XTI/TLI and socket-based interfaces before you decide to port an application.

The following issues are related to transport independence, and can have some bearing on RPC applications:

- *Privileged ports* – Privileged ports are an artifact of the Berkeley Software Distribution (BSD) implementation of the TCP/IP Internet Protocols. They are not portable. The notion of privileged ports is not supported in the transport-independent environment.
- *Opaque addresses* – There is no transport-independent way of separating the portion of an address that names a host from the portion of an address that names the service at that host. Be sure to change any code that assumes it can discern the host address of a network service.
- *Broadcast* – There is no transport-independent form of broadcast address.

Socket-to-XTI/TLI Equivalents

Table 3–15 shows approximate equivalents between XTI/TLI functions and socket functions. The comment field describes the differences. If there is no comment, either the functions are similar or there is no equivalent function in either interface.

TABLE 3–15 TLI and Socket Equivalent Functions

TLI function	Socket function	Comments
`t_open`(3NSL)	`socket`(3SOCKET)	
--	`socketpair`(3SOCKET)	

TABLE 3–15 TLI and Socket Equivalent Functions *(continued)*

TLI function	Socket function	Comments
t_bind(3NSL)	bind(3SOCKET)	t_bind(3NSL) sets the queue depth for passive sockets, but bind(3SOCKET) doesn't. For sockets, the queue length is specified in the call to listen(3SOCKET).
t_optmgmt(3NSL)	getsockopt(3SOCKET) setsockopt(3SOCKET)	t_optmgmt(3NSL) manages only transport options. getsockopt(3SOCKET) and setsockopt(3SOCKET) can manage options at the transport layer, but also at the socket layer and at the arbitrary protocol layer.
t_unbind(3NSL)		--
t_close(3NSL)	close(2)	
t_getinfo(3NSL)	getsockopt(3SOCKET)	t_getinfo(3NSL) returns information about the transport. getsockopt(3SOCKET) can return information about the transport and the socket.
t_getstate(3NSL)	-	
t_sync(3NSL)	-	
t_alloc(3NSL)	-	
t_free(3NSL)	-	
t_look(3NSL)	-	getsockopt(3SOCKET) with the SO_ERROR option returns the same kind of error information as t_look(3NSL)t_look( ).
t_error(3NSL)	perror(3C)	

TABLE 3–15 TLI and Socket Equivalent Functions *(continued)*

TLI function	Socket function	Comments
t_connect(3NSL)	connect(3SOCKET)	A connect(3SOCKET) can be done without first binding the local endpoint. The endpoint must be bound before calling t_connect(3NSL). A connect(3SOCKET) can be done on a connectionless endpoint to set the default destination address for datagrams. Data can be sent on a connect(3SOCKET).
t_rcvconnect(3NSL)	-	
t_listen(3NSL)	listen(3SOCKET)	t_listen(3NSL) waits for connection indications. listen(3SOCKET) merely sets the queue depth.
t_accept(3NSL)	accept(3SOCKET)	
t_snd(3NSL)	send(3SOCKET)	
	sendto(3SOCKET)	
	sendmsg(3SOCKET)	sendto(3SOCKET) and sendmsg(3SOCKET) operate in connection mode as well as datagram mode.
t_rcv(3NSL)	recv(3SOCKET)	
	recvfrom(3SOCKET)	
	recvmsg(3SOCKET)	recvfrom(3SOCKET) and recvmsg(3SOCKET) operate in connection mode as well as datagram mode.
t_snddis(3NSL)	-	
t_rcvdis(3NSL)	-	

TABLE 3–15 TLI and Socket Equivalent Functions *(continued)*

TLI function	Socket function	Comments
t_sndrel(3NSL)	shutdown(3SOCKET)	
t_rcvrel(3NSL)	-	
t_sndudata(3NSL)	sendto(3SOCKET)	
	recvmsg(3SOCKET)	
t_rcvuderr(3NSL)	-	
read(2), write(2)	read(2), write(2)	In XTI/TLI you must push the tirdwr(7M) module before calling read(2) or write(2); in sockets, just call read(2) or write(2).

Additions to XTI Interface

The XNS 5 (Unix98) standard introduces some new XTI interfaces. These are briefly described below. The details may be found in the relevant manual pages. These interfaces are not available for TLI users.

Scatter/Gather Data Transfer Interfaces

t_sndvudata**(3NSL)**	Send a data unit from one or more non-contiguous buffers
t_rcvvudata**(3NSL)**	Receive a data unit into one or more non-contiguous buffers

`t_sndv`**(3NSL)**	Send data or expedited data from one or more non-contiguous buffers on a connection
`t_rcvv`**(3NSL)**	Receive data or expedited data sent over a connection and put the data into one or more non-contiguous buffers

XTI Utility Functions

`t_sysconf`**(3NSL)**	Get configurable XTI variables

Additional Connection Release Interfaces

`t_sndreldata`**(3NSL)**	Initiate / respond to an orderly release with user data
`t_rcvreldata`**(3NSL)**	Receive an orderly release indication or confirmation containing user data

---

Note - The additional interfaces `t_sndreldata`(3NSL) and `t_rcvreldata`(3NSL) are only for use with a specific transport called “minimal OSI”, which is not available on the Solaris platform. These interfaces are not available for use in conjunction with Internet Transports (TCP or UDP).

CHAPTER 4

Transport Selection and Name-to-Address Mapping

This chapter describes selecting transports and resolving network addresses. It further describes interfaces that enable you to specify the available communication protocols for an application. The chapter also explains additional functions that provide direct mapping of names to network addresses.

Note - In this chapter the terms *network* and *transport* are used interchangeably to refer to the programmatic interface that conforms to the transport layer of the OSI Reference Mode. The term network is also used to refer to the physical collection of computers connected through some electronic medium.

Transport Selection Is Multithread Safe

The interface described in this chapter is multithread safe. This means that applications that contain transport selection function calls can be used freely in a multithreaded application. Note, however, that the degree of concurrency available to applications is not specified.

Transport Selection

A distributed application must use a standard interface to the transport services to be portable to different protocols. Transport selection services provide an interface that allows an application to select which protocols to use. This makes an application "protocol" and "medium" independent.

Transport selection makes it easy for a client application to try each available transport until it establishes communication with a server. Transport selection lets server applications accept requests on multiple transports, and in doing so, communicate over a number of protocols. Transports can be tried in either the order specified by the local default sequence or in an order specified by the user.

Choosing from the available transports is the responsibility of the application. The transport selection mechanism makes that selection uniform and simple.

How Transport Selection Works

The transport selection component is built around:

- A network configuration database (the `/etc/netconfig` file), which contains an entry for each network on the system
- Optional use of the `NETPATH` environment variable

The `NETPATH` variable is set by the user; it contains an ordered list of transport identifiers. The transport identifiers match the `netconfig` network ID field and are links to records in the `netconfig`(4) file. The `netconfig`(4) file is described in "`/etc/netconfig` File" on page 121. The network selection interface is a set of access routines for the network-configuration database.

One set of library routines accesses only the `/etc/netconfig` entries identified by the `NETPATH` environment variable:

`setnetpath`(3NSL)	Initializes the search of `NETPATH`
`getnetpath`(3NSL)	Returns a pointer to the `netconfig`(4) entry that corresponds to the next component of the `NETPATH` variable
`endnetpath`(3NSL)	Releases the database pointer to elements in the `NETPATH` variable when processing is complete

These routines are described in "`NETPATH` Access to `netconfig`(4) Data" on page 124 and in `getnetpath`(3NSL). They let the user influence the selection of transports used by the application.

To avoid user influence on transport selection, use the routines that access the netconfig(4) database directly. These routines are described in "Accessing netconfig(4)" on page 125 and in getnetconfig(3NSL):

setnetconfig(3NSL)	Initializes the record pointer to the first index in the database
getnetconfig(3NSL)	Returns a pointer to the current record in the netconfig(4) database and increments the pointer to the next record
endnetconfig(3NSL)	Releases the database pointer when processing is complete

The following two routines manipulate netconfig(4) entries and the data structures they represent. These routines are described in "Accessing netconfig(4)" on page 125:

getnetconfigent(3NSL)	Returns a pointer to the struct netconfig structure corresponding to netid
freenetconfigent(3NSL)	Frees the structure returned by getnetconfigent(3NSL)

/etc/netconfig File

The netconfig(4) file describes all transport protocols on a host. The entries in the netconfig(4) file are explained briefly in Table 4–1 and in more detail in the netconfig(4) man page.

TABLE 4–1 netconfig(4) File

Entries	Description
network ID	A local representation of a transport name (such as tcp). Do not assume that this field contains a well-known name (such as tcp or udp) or that two systems use the same name for the same transport.
semantics	The semantics of the particular transport protocol. Valid semantics are: ■ tpi_clts – connectionless ■ tpi_cots – connection oriented ■ tpi_cots_ord – connection oriented with orderly release
flags	Can take only the values, v, or hyphen (-). Only the visible flag (-v) is defined.
protocol family	The protocol family name of the transport provider (for example, inet or loopback).

TABLE 4–1 netconfig(4) File *(continued)*

Entries	Description
protocol name	The protocol name of the transport provider. For example, if *protocol family* is inet, then *protocol name* is tcp, udp, or icmp. Otherwise, the value of *protocol name* is a hyphen (-).
network device	The full path name of the device file to open when accessing the transport provider
name-to-address translation libraries	Names of the shared objects. This field contains the comma-separated file names of the shared objects that contain name-to-address mapping routines. Shared objects are located through the path in the LD_LIBRARY_PATH variable. A "-" in this field indicates redirection to the name service switch policies for hosts and services.

Code Example 4–1 shows a sample netconfig(4) file. Use of the netconfig(4) file has been changed for the inet transports, as described in the commented section in the sample file. This change is also described in "Name-to-Address Mapping" on page 127.

CODE EXAMPLE 4–1 Sample netconfig(4) File

```
# The ‘‘Network Configuration’’ File.
#
# Each entry is of the form:
#
#<net <semantics>  <flags> <proto    <proto   <device>     <nametoaddr_libs>
# id>                       family>   name>
#
# The "-" in <nametoaddr_libs> for inet family transports indicates redirection
# to the name service switch policies for "hosts" and "services. The "-" may be
# replaced by nametoaddr libraries that comply with the SVR4 specs, in which
# case the name service switch will be used for netdir_getbyname, netdir_
# getbyaddr, gethostbyname, gethostbyaddr, getservbyname, and getservbyport.
# There are no nametoaddr_libs for the inet family in Solaris anymore.
#
udp       tpi_clts     v   inet       udp      /dev/udp        -
#
tcp       tpi_cots_ord v   inet       tcp      /dev/tcp        -
#
icmp      tpi_raw      -   inet       icmp     /dev/icmp       -
#
rawip     tpi_raw      -   inet       -        /dev/rawip      -
#
ticlts    tpi_clts     v   loopback   -        /dev/ticlts     straddr.so
#
ticots    tpi_cots     v   loopback   -        /dev/ticots     straddr.so
#
ticotsord tpi_cots_ord v   loopback   -        /dev/ticotsord  straddr.so
#
```

Network selection library routines return pointers to `netconfig` entries. The `netconfig` structure is shown in Code Example 4–2.

CODE EXAMPLE 4–2 `netconfig` Structure

```
struct netconfig {
   char  *nc_netid;                /* network identifier */
   unsigned  int   nc_semantics;   /* semantics of protocol */
   unsigned  int   nc_flag;        /* flags for the protocol */
   char *nc_protofmly;             /* family name */
   char *nc_proto;                 /* proto specific */
   char *nc_device;                /* device name for network id */
   unsigned  int   nc_nlookups;    /* # entries in nc_lookups */
   char **nc_lookups;              /* list of lookup libraries */
   unsigned  int   nc_unused[8];
};
```

Valid network IDs are defined by the system administrator, who must ensure that network IDs are locally unique. If they are not, some network selection routines can fail. For example, it is not possible to know which network `getnetconfigent("udp")` will use if there are two `netconfig` entries with the network ID `udp`.

The system administrator also sets the order of the entries in the `netconfig`(4) database. The routines that find entries in `/etc/netconfig` return them in order, from the beginning of the file. The order of transports in the `netconfig`(4) file is the default transport search sequence of the routines. Loopback entries should be at the end of the file.

The `netconfig`(4) file and the `netconfig` structure are described in greater detail in the `netconfig`(4) man page.

`NETPATH` Environment Variable

An application usually uses the default transport search path set by the system administrator to locate an available transport. However, when a user wants to influence the choices made by an application, the application can modify the interface by using the environment variable `NETPATH` and the routines described in the section, “`NETPATH` Access to `netconfig`(4) Data” on page 124. These routines access only the transports specified in the `NETPATH` variable.

`NETPATH` is similar to the `PATH` variable. It is a colon-separated list of transport IDs. Each transport ID in the `NETPATH` variable corresponds to the network ID field of a record in the `netconfig`(4) file. `NETPATH` is described in the `environ`(4) man page.

The default transport set is different for the routines that access `netconfig`(4) through the `NETPATH` environment variable (described in the next section) and the routines that access `netconfig`(4) directly. The default transport set for routines that access `netconfig`(4) via `NETPATH` consists of the visible transports in the

netconfig(4) file. For routines that access netconfig(4) directly, the default transport set is the entire netconfig(4) file. A transport is visible if the system administrator has included a v flag in the flags field of that transport's netconfig(4) entry.

NETPATH Access to netconfig(4) Data

Three routines access the network configuration database indirectly through the NETPATH environment variable. The variable specifies the transport or transports an application is to use and the order to try them. NETPATH components are read from left to right. The functions have the following interfaces:

```
#include <netconfig.h>

void *setnetpath(void);
struct netconfig *getnetpath(void *);
int endnetpath(void *);
```

A call to setnetpath(3NSL) initializes the search of NETPATH. It returns a pointer to a database that contains the entries specified in a NETPATH variable. The pointer, called a handle, is used to traverse this database with getnetpath(3NSL). The setnetpath(3NSL) function must be called before the first call to getnetpath(3NSL).

When first called, getnetpath(3NSL) returns a pointer to the netconfig(4) file entry that corresponds to the first component of the NETPATH variable. On each subsequent call, getnetpath(3NSL) returns a pointer to the netconfig(4) entry that corresponds to the next component of the NETPATH variable; getnetpath(3NSL) returns NULL if there are no more components in NETPATH. A call to getnetpath(3NSL) without an initial call to setnetpath(3NSL) causes an error; getnetpath(3NSL) requires the pointer returned by setnetpath(3NSL) as an argument.

getnetpath(3NSL) silently ignores invalid NETPATH components. A NETPATH component is invalid if there is no corresponding entry in the netconfig(4) database.

If the NETPATH variable is unset, getnetpath(3NSL) behaves as if NETPATH were set to the sequence of default or visible transports in the netconfig(4) database, in the order in which they are listed.

endnetpath(3NSL) is called to release the database pointer to elements in the NETPATH variable when processing is complete. endnetpath(3NSL) fails if setnetpath(3NSL) was not called previously. Code Example 4–3 shows the setnetpath(3NSL), getnetpath(3NSL), and endnetpath(3NSL) routines.

CODE EXAMPLE 4–3 `setnetpath`(3NSL), `getnetpath`(3NSL), and `endnetpath`(3NSL) Functions

```
#include <netconfig.h>

void *handlep;
struct netconfig *nconf;

if ((handlep = setnetpath()) == (void *)NULL) {
   nc_perror(argv[0]);
   exit(1);
}

while ((nconf = getnetpath(handlep)) != (struct netconfig *)NULL)
{
   /*
    * nconf now describes a transport provider.
    */
}
endnetpath(handlep);
```

The `netconfig`(4) structures obtained through `getnetpath`(3NSL) become invalid after the execution of `endnetpath`(3NSL). To preserve the data in the structure, use `getnetconfigent(nconf->nc_netid)` to copy them into a new data structure.

Accessing `netconfig`(4)

Three functions access `/etc/netconfig` and locate `netconfig`(4) entries. The routines `setnetconfig`(3NSL), `getnetconfigent`(3NSL), and `endnetconfig`(3NSL) have the following interfaces:

```
#include <netconfig.h>

void *setnetconfig(void);
struct netconfig *getnetconfig(void *);
int endnetconfig(void *);
```

A call to `setnetconfig`(3NSL) initializes the record pointer to the first index in the database; `setnetconfig`(3NSL) must be used before the first use of `getnetconfig`(3NSL). `setnetconfig`(3NSL) returns a unique handle (a pointer into the database) to be used by the `getnetconfig`(3NSL) routine. Each call to `getnetconfig`(3NSL) returns the pointer to the current record in the `netconfig`(4) database and increments its pointer to the next record. It can be used to search the entire `netconfig`(4) database. `getnetconfig`(3NSL) returns a `NULL` at the end of file.

You must use `endnetconfig`(3NSL) to release the database pointer when processing is complete. `endnetconfig`(3NSL) must not be called before `setnetconfig`(3NSL).

CODE EXAMPLE 4–4 `setnetconfig`(3NSL), `getnetconfig`(3NSL), and `endnetconfig`(3NSL) Functions

```
void *handlep;
struct netconfig *nconf;

if ((handlep = setnetconfig()) == (void *)NULL){
   nc_perror(argv[0]);
   exit(1);
}
/*
 * transport provider information is described in nconf.
 * process_transport is a user-supplied routine that
 * tries to connect to a server over transport nconf.
 */
while ((nconf = getnetconfig(handlep)) != (struct netconfig *)NULL){
    if (process_transport(nconf) == SUCCESS)
      break;
}
endnetconfig(handlep);
```

The last two functions have the following interface:

```
#include <netconfig.h>
struct netconfig *getnetconfigent(char *);
int freenetconfigent(struct netconfig *);
```

`getnetconfigent`(3NSL) returns a pointer to the `struct netconfig` structure corresponding to `netid`. It returns `NULL` if `netid` is invalid. `setnetconfig`(3NSL) need not be called before `getnetconfigent`(3NSL).

`freenetconfigent`(3NSL) frees the structure returned by `getnetconfigent`(3NSL). Code Example 4–5 shows the `getnetconfigent`(3NSL) and `freenetconfigent`(3NSL) routines.

CODE EXAMPLE 4–5 `getnetconfigent`(3NSL) and `freenetconfigent`(3NSL) Functions

```
/* assume udp is a netid on this host */
struct netconfig *nconf;

if ((nconf = getnetconfigent(''udp'')) == (struct netconfig *)NULL){
   nc_perror(''no information about udp'');
   exit(1);
}
process_transport(nconf);
freenetconfigent(nconf);
```

Loop Through All Visible `netconfig`(4) Entries

The `setnetconfig`(3NSL) call is used to step through all the transports marked *visible* (by a `v` flag in the `flags` field) in the `netconfig`(4) database. The transport selection routine returns a `netconfig`(4) pointer.

Looping Through User-Defined netconfig(4) Entries

Users can control the loop by setting the NETPATH environment variable to a colon-separated list of transport names. If NETPATH is set as follows:

```
NETPATH=tcp:udp
```

The loop first returns the tcp entry, then the udp entry. If NETPATH is not defined, the loop returns all visible entries in the netconfig(4) file in the order in which they are stored. The NETPATH environment variable lets users define the order in which client-side applications try to connect to a service. It also lets the server administrator limit transports on which a service can listen.

Use getnetpath(3NSL) and setnetpath(3NSL) to obtain or modify the network path variable. Code Example 4–6 shows the form and use, which are similar to the getnetconfig(3NSL) and setnetconfig(3NSL) routines.

CODE EXAMPLE 4–6 Looping Through Visible Transports

```
void *handlep;
struct netconfig *nconf;

if ((handlep = setnetconfig() == (void *) NULL) {
    nc_perror(‘‘setnetconfig’’);
   exit(1);
}
while (nconf = getnetconfig(handlep))
    if (nconf->nc_flag & NC_VISIBLE)
      doit(nconf);
(void) endnetconfig(handlep);
```

Name-to-Address Mapping

Name-to-address mapping lets an application obtain the address of a service on a specified host, independent of the transport used. Name-to-address mapping consists of the following functions:

netdir_getbyname(3NSL)	Maps the host and service name to a set of addresses
netdir_getbyaddr(3NSL)	Maps addresses into host and service names
netdir_free(3NSL)	Frees structures allocated by the name-to-address translation routines

`taddr2uaddr`(3NSL)	Translates an address and returns a transport-independent character representation of the address
`uaddr2taddr`(3NSL)	The universal address is translated into a `netbuf` structure
`netdir_options`(3NSL)	Interfaces to transport-specific capabilities (such as the broadcast address and reserved port facilities of TCP and UDP)

The first argument of each routine points to a `netconfig`(4) structure that describes a transport. The routine uses the array of directory-lookup library paths in the `netconfig`(4) structure to call each path until the translation succeeds.

The libraries are described in Table 4–2. The routines described in the section, “Using the Name-to-Address Mapping Routines” on page 129, are defined in the `netdir`(3NSL) man page.

Note - The following libraries no longer exist in the Solaris 2 environment: `tcpip.so`, `switch.so`, and `nis.so`. For more information on this change, see the `nsswitch.conf`(4) man page and the NOTES section of the `gethostbyname`(3NSL) man page.

TABLE 4–2 Name-to-Address Libraries

Library	Transport Family	Description
-	`inet`	For networks of the protocol family `inet`, its name-to-address mapping is provided by the name service switch based on the entries for *hosts* and *services* in the file `nsswitch.conf`(4). For networks of other families, the "-" indicates a nonfunctional name-to-address mapping.
`straddr.so`	`loopback`	Contains the name-to-address mapping routines of any protocol that accepts strings as addresses, such as the loopback transports.

`straddr.so` Library

Name–to–address translation files for the library are created and maintained by the system administrator. The `straddr.so` files are `/etc/net/`*transport-name*`/hosts` and `/etc/net/`*transport-name*`/services`. *transport-name* is the local name of the transport that accepts string addresses (specified in the *network ID* field of the `/etc/`

netconfig file). For example, the host file for ticlts would be /etc/net/ticlts/hosts, and the service file for ticlts would be /etc/net/ticlts/services.

Even though most string addresses do not distinguish between *host* and *service*, separating the string into a host part and a service part is consistent with other transports. The /etc/net/*transport-name*/hosts file contains a text string that is assumed to be the host address, followed by the host name. For example:

```
joyluckaddr joyluck
carpediemaddr carpediem
thehopaddr thehop
pongoaddr pongo
```

For loopback transports, it makes no sense to list other hosts because the service cannot go outside the containing host.

The /etc/net/transport-name/services file contains service names followed by strings identifying the service address. For example:

```
rpcbind rpc
listen serve
```

The routines create the full-string address by concatenating the host address, a period (.), and the service address. For example, the address of the listen service on pongo is pongoaddr.serve.

When an application requests the address of a service on a particular host on a transport that uses this library, the host name must be in /etc/net/*transport*/hosts, and the service name must be in /etc/net/*transport*/services. If either is missing, the name-to-address translation fails.

Using the Name-to-Address Mapping Routines

This section is an overview of what routines are available to use. The routines return or convert the network names to their respective network addresses. Note that netdir_getbyname(3NSL), netdir_getbyaddr(3NSL), and taddr2uaddr(3NSL) return pointers to data that must be freed by calls to netdir_free(3NSL).

```
int netdir_getbyname(struct netconfig *nconf,
  struct nd_hostserv *service, struct nd_addrlist **addrs);
```

netdir_getbyname(3NSL) maps the host and service name specified in *service* to a set of addresses consistent with the transport identified in *nconf*. The nd_hostserv and nd_addrlist structures are defined in the netdir(3NSL) man page. A pointer to the addresses is returned in *addrs*.

To find all addresses of a host and service (on all available transports), call `netdir_getbyname`(3NSL) with each `netconfig`(4) structure returned by either `getnetpath`(3NSL) or `getnetconfig`(3NSL).

```
int netdir_getbyaddr(struct netconfig *nconf,
  struct nd_hostservlist **service, struct netbuf *netaddr);
```

`netdir_getbyaddr`(3NSL) maps addresses into host and service names. The function is called with an address in *netaddr* and returns a list of host-name and service-name pairs in *service*. The `nd_hostservlist` structure is defined in `netdir`(3NSL).

```
void netdir_free(void *ptr, int struct_type);
```

The `netdir_free`(3NSL) routine frees structures allocated by the name-to-address translation routines. The parameters can take the values shown in Table 4–3.

TABLE 4–3 `netdir_free`(3NSL) Routines

struct_type	ptr
`ND_HOSTSERV`	Pointer to an `nd_hostserv` structure
`ND_HOSTSERVLIST`	Pointer to an `nd_hostservlist` structure
`ND_ADDR`	Pointer to a `netbuf` structure
`ND_ADDRLIST`	Pointer to an `nd_addrlist` structure

```
char *taddr2uaddr(struct netconfig *nconf, struct netbuf *addr);
```

`taddr2uaddr`(3NSL) translates the address pointed to by *addr* and returns a transport-independent character representation of the address (“universal address”). *nconf* specifies the transport for which the address is valid. The universal address can be freed by `free`(3C).

```
struct netbuf *uaddr2taddr(struct netconfig *nconf, char *uaddr);
```

The “universal address” pointed to by *uaddr* is translated into a `netbuf` structure; *nconf* specifies the transport for which the address is valid.

```
int netdir_options(struct netconfig *nconf,
   int dt, int dt,
   char *point_to_args);
```

`netdir_options`(3NSL) interfaces to transport-specific capabilities (such as the broadcast address and reserved port facilities of TCP and UDP). *nconf* specifies a transport. *option* specifies the transport-specific action to take. *fd* might or might not be used depending upon the value of *option*. The fourth argument points to operation-specific data.

Table 4–4 shows the values used for `option`:

TABLE 4–4 Values for `netdir_options`

Option	Description
`ND_SET_BROADCAST`	Sets the transport for broadcast (if the transport supports broadcast)
`ND_SET_RESERVEDPORT`	Lets the application bind to a reserved port (if allowed by the transport)
`ND_CHECK_RESERVEDPORT`	Verifies that an address corresponds to a reserved port (if the transport supports reserved ports)
`ND_MERGEADDR`	Transforms a locally meaningful address into an address to which client hosts can connect

`netdir_perror`(3NSL) displays the message stating why one of the name-to-address mapping routines failed on `stderr`.

```
void netdir_perror(char *s);
```

`netdir_sperror`(3NSL) returns a string containing the error message stating why one of the name-to-address mapping routines failed.

```
char *netdir_sperror(void);
```

Code Example 4–7 shows network selection and name-to-address mapping.

CODE EXAMPLE 4–7 Network Selection and Name-to-Address Mapping

```
#include <netconfig.h>
#include <netdir.h>
#include <sys/tiuser.h>

struct nd_hostserv nd_hostserv;   /* host and service information */
struct nd_addrlist *nd_addrlistp; /* addresses for the service */
struct netbuf *netbufp;           /* the address of the service */
struct netconfig *nconf;          /* transport information*/
int i;                            /* the number of addresses */
char *uaddr;                      /* service universal address */
void *handlep;                    /* a handle into network selection */
/*
```

```
 * Set the host structure to reference the "date"
 * service on host "gandalf"
 */
nd_hostserv.h_host = "gandalf";
nd_hostserv.h_serv = "date";
/*
 * Initialize the network selection mechanism.
 */
if ((handlep = setnetpath()) == (void *)NULL) {
   nc_perror(argv[0]);
   exit(1);
}
/*
 * Loop through the transport providers.
 */
while ((nconf = getnetpath(handlep)) != (struct netconfig *)NULL)
{
   /*
    * Print out the information associated with the
    * transport provider described in the "netconfig"
    * structure.
    */
   printf("Transport provider name: %s\n", nconf->nc_netid);
   printf("Transport protocol family: %s\n", nconf->nc_protofmly);
   printf("The transport device file: %s\n", nconf->nc_device);
   printf("Transport provider semantics: ");
    switch (nconf->nc_semantics) {
   case NC_TPI_COTS:
      printf("virtual circuit\n");
      break;
   case NC_TPI_COTS_ORD:
      printf("virtual circuit with orderly release\n");
      break;

   case NC_TPI_CLTS:
      printf("datagram\n");
      break;
   }
   /*
    * Get the address for service "date" on the host
    * named "gandalf" over the transport provider
    * specified in the netconfig structure.
    */
   if (netdir_getbyname(nconf, &nd_hostserv, &nd_addrlistp) != ND_OK) {
      printf("Cannot determine address for service\n");
      netdir_perror(argv[0]);
      continue;
   }
   printf("<%d> addresses of date service on gandalf:\n",
      nd_addrlistp->n_cnt);
   /*
    * Print out all addresses for service "date" on
    * host "gandalf" on current transport provider.
    */
   netbufp = nd_addrlistp->n_addrs;
   for (i = 0; i < nd_addrlistp->n_cnt; i++, netbufp++) {
      uaddr = taddr2uaddr(nconf,netbufp);
      printf("%s\n",uaddr);
      free(uaddr);
   }
```

```
    netdir_free( nd_addrlistp, ND_ADDRLIST );

}
endnetconfig(handlep);
```

APPENDIX **A**

UNIX Domain Sockets

Introduction

UNIX domain sockets are named with UNIX paths. For example, a socket might be named `/tmp/foo`. UNIX domain sockets communicate only between processes on a single host. Sockets in the UNIX domain are not considered part of the network protocols because they can only be used to communicate between processes on a single host.

Socket types define the communication properties visible to a user. The Internet domain sockets provide access to the TCP/IP transport protocols. The Internet domain is identified by the value `AF_INET`. Sockets exchange data only with sockets in the same domain.

Socket Creation

The `socket`(3SOCKET) call creates a socket in the specified family and of the specified type.

```
s = socket(family, type, protocol);
```

If the protocol is unspecified (a value of `0`), the system selects a protocol that supports the requested socket type. The socket handle (a file descriptor) is returned.

The family is specified by one of the constants defined in `sys/socket.h`. Constants named `AF_`*suite* specify the address format to use in interpreting names as shown in Table 2–1.

The following creates a datagram socket for intramachine use:

```
s = socket(AF_UNIX, SOCK_DGRAM, 0);
```

Use the default protocol (the *protocol* argument is `0`) in most situations.

Binding Local Names

A socket is created with no name. A remote process has no way to refer to a socket until an address is bound to it. Communicating processes are connected through addresses. In the UNIX family, a connection is composed of (usually) one or two path names. UNIX family sockets need not always be bound to a name, but, when bound, there can never be duplicate ordered sets such as: `local pathname` or `foreign pathname`. The path names cannot refer to existing files.

The `bind`(3SOCKET) call allows a process to specify the local address of the socket. This provides `local pathname`, while `connect`(3SOCKET) and `accept`(3SOCKET) complete a socket's association by fixing the remote half of the address. `bind`(3SOCKET) is used as follows:

```
bind (s, name, namelen);
```

The socket handle is *s*. The bound name is a byte string that is interpreted by the supporting protocol(s). UNIX family names contain a path name and a family. The example shows binding the name `/tmp/foo` to a UNIX family socket.

```
#include <sys/un.h>
 ...
struct sockaddr_un addr;
 ...
strcpy(addr.sun_path, "/tmp/foo");
addr.sun_family = AF_UNIX;
bind (s, (struct sockaddr *) &addr,
  strlen(addr.sun_path) + sizeof (addr.sun_family));
```

When determining the size of an `AF_UNIX` socket address, null bytes are not counted, which is why `strlen`(3C) use is fine.

The file name referred to in `addr.sun_path` is created as a socket in the system file name space. The caller must have write permission in the directory where `addr.sun_path` is created. The file should be deleted by the caller when it is no longer needed. `AF_UNIX` sockets can be deleted with `unlink`(1M).

Connection Establishment

Connection establishment is usually asymmetric. One process acts as the client and the other as the server. The server binds a socket to a well-known address associated with the service and blocks on its socket for a connect request. An unrelated process can then connect to the server. The client requests services from the server by initiating a connection to the server's socket. On the client side, the

connect(3SOCKET) call initiates a connection. In the UNIX family, this might appear as:

```
struct sockaddr_un server;
  server.sun.family = AF_UNIX;
   ...
  connect(s, (struct sockaddr *)&server, strlen(server.sun_path)
         + sizeof (server.sun_family));
```

See “Connection Errors” on page 22 for information on connection errors. “Data Transfer” on page 23 tells you how to transfer data. “Closing Sockets” on page 24 tells you how to close a socket.

APPENDIX **B**

Live Code Example

Live Code Examples

What follows in this appendix are copies of the complete live code modules used in this book. They are compilable as they sit and will run (unless otherwise noted to be pseudo-code or the like). They are provided for informational purposes only and Solaris Software assumes no liability from their use.

Index

U

X

Z