Patricia Morreale et al. "Voice and Data Communications"
The CRC Handbook of Modern Telecommunications
Ed. Patricia Morreale and Kornel Terplan
Boca Raton, CRC Press LLC. 2001

1

Voice and Data Communications

Patricia Morreale
Stevens Institute of Technology

Michel Gilbert
Hill Associates, Inc.

Matthew Kolon
Hill Associates, Inc.

John Amoss
Lucent Technologies

1.1 Advanced Intelligent Networks (AIN)

Patricia Morreale

1.1.1 Definition

Intelligent network (IN) is a telephone network architecture originated by Bell Communications Research (Bellcore) in which the service logic for a call is located separately from the switching facilities, allowing services to be added or changed without having to redesign switching equipment. According to Bell Atlantic, IN is a service-specific architecture. That is, a certain portion of a dialed phone number, such as 800 or 900, triggers a request for a specific service. A later version of IN called advanced intelligent network (AIN) introduces the idea of a service-independent architecture in which a given part of a telephone number can be interpreted differently by various services depending on factors such as time of day, caller identity, and type of call. AIN makes it easy to add new services without having to install new phone equipment.

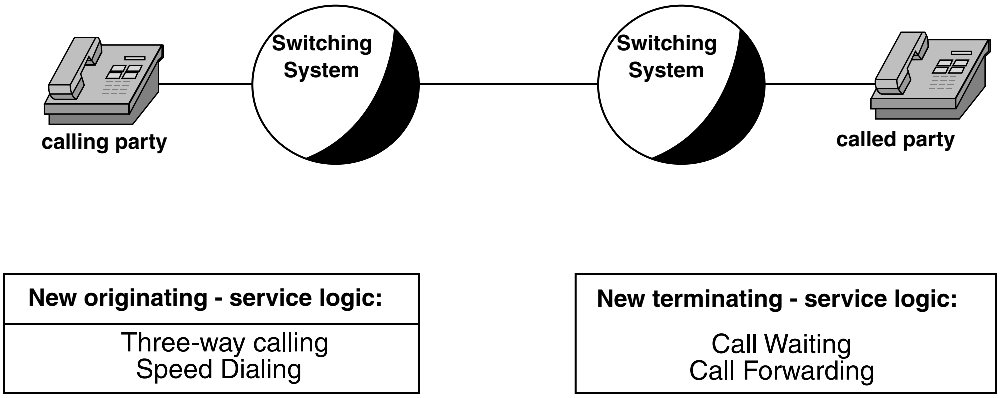

FIGURE 1.1.1 Plain old telephone service (POTS).

1.1.2 Overview

This chapter discusses how the network has evolved from one in which switch-based service logic provides services to one in which service-independent AIN capabilities allow for service creation and deployment.

As the IN evolves, service providers will be faced with many opportunities and challenges. While the IN provides a network capability to meet the ever-changing needs of customers, network intelligence is becoming increasingly distributed and complicated. For example, third-party service providers will be interconnecting with traditional operating company networks. Local number portability (LNP) presents many issues that can only be resolved in an IN environment to meet government mandates. Also, as competition grows with companies offering telephone services previously denied to them, the IN provides a solution to meet the challenge.

1.1.3 Network Evolution

1.1.3.1 Plain Old Telephone Service (POTS)

Prior to the mid-1960s, the service logic (Figure 1.1.1) was hard-wired in switching systems. Typically, network operators met with switch vendors, discussed the types of services customers required, negotiated the switching features that provided the services, and finally agreed upon a generic release date for feature availability. After this, the network operator planned for the deployment of the generic feature/service in the switching network fabric.

This process was compounded for the network operator with switching systems from multiple vendors. As a result, services were not offered ubiquitously across an operator's serving area. So, a customer in one end of a city, county, or state may not have had the same service offerings as a person in another part of the area.

Also, once services were implemented, they were not easily modified to meet individual customer's requirements. Often, the network operator negotiated the change with the switch vendor. As a result of this process, it took years to plan and implement services. This approach to new service deployment required detailed management of calling patterns, and providing new trunk groups to handle calling patterns. As customer calling habits changed — such as longer call lengths, larger calling areas, and multiple lines in businesses and residences — the demand on network operators increased.

1.1.3.2 Stored Program Control (SPC)

In the mid-1960s, stored program control (SPC) switching systems were introduced. SPC was a major step forward because now service logic was programmable where, in the past, the service logic was hard wired. As a result, it was now easier to introduce new services. Nevertheless, this service logic concept was not modular. It became increasingly more complicated to add new services because of the dependency between the service and the service-specific logic. Essentially, service logic that was used for one service

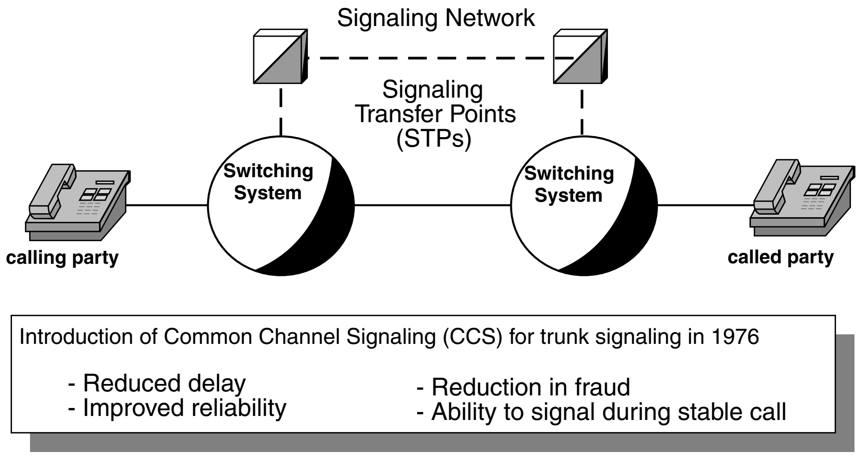

FIGURE 1.1.2 Common channel signaling (CCS).

could not be used for another. As a result, if customers were not served by a SPC switching system, new services were not available to them.

1.1.3.3 Common Channel Signaling Network (CCSN)

Another aspect of the traditional service offerings was the call setup information — the signaling and call supervision that took place between switching systems and the actual call. When a call was set up, a signal and talk path used the same common trunk from the originating switching system to the terminating switching system. Often there were multiple offices involved in the routing of a call. This process seized the trunks in all of the switching systems involved. Hence, if the terminating end was busy, all of the trunks were set up unnecessarily.

The network took a major leap forward in the mid-1970s with the introduction of the common channel signaling network (CCSN), or SS7 network for short. Signaling system number 7 (SS7) is the protocol that runs over the CCSN. The SS7 network consists of packet data links and packet data switching systems called signaling transfer points (STPs).

The SS7 network (Figure 1.1.2) separates the call setup information and talk path from the common trunks that run between switching systems. The call setup information travels outside the common trunk path over the SS7 network. The type of information transferred includes permission for the call setup, whether or not the called party is busy.

SS7 technology frees up trunk circuits between switching systems for the actual calls. The SS7 network enabled the introduction of new services, such as caller ID. Caller ID provides the calling party's telephone number, which is transmitted over the SS7 network. The SS7 network was designed before the IN concept was introduced. However, telephone operators realized that there were many advantages to implementing and using SS7 network capabilities.

1.1.4 Introduction of IN

During the mid-1980s, regional Bell operating companies (RBOCs) began requesting features that met the following objectives:

- Rapid deployment of services in the network
- Vendor independence and standard interfaces
- Opportunities for non-RBOCs to offer services for increased network usage

Bell Communications Research (Bellcore) responded to this request and developed the concept of Intelligent Network 1 (IN/1, Figure 1.1.3).

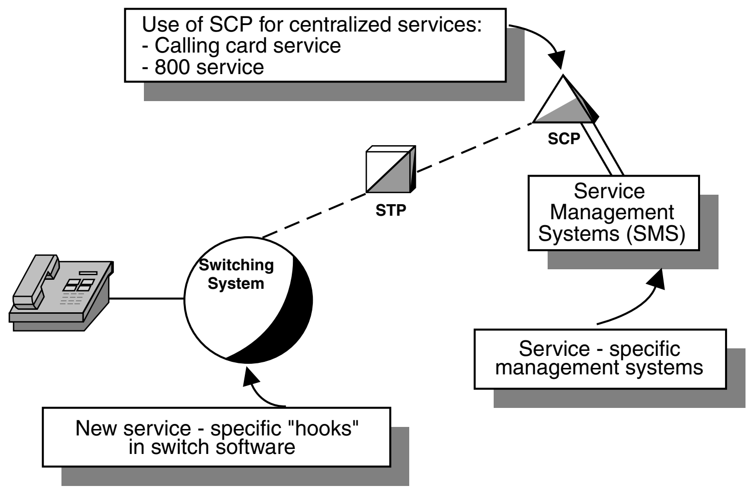

FIGURE 1.1.3 Intelligent Network (IN/1).

The introduction of the IN/1 marked the first time that service logic was external to switching systems and located in databases called service control points (SCPs). Two services evolved that required IN/1 service logic — the 800 (or Freephone) service and the calling card verification (or alternate billing service, ABS) service. Because of the service-specific nature of the technology, these services required two separate SCPs. In order to communicate with the associated service logic, software was deployed in switching systems. This switching system software enabled the switching system to recognize when it was necessary to communicate with a SCP via the SS7 network. With the introduction of the SCP concept, new operations and management systems became necessary to support service creation, testing, and provisioning. In Figure 1.1.3, note the term "service-specific management systems" under the box labeled "service management system." This means that the software-defined "hooks" or triggers are specific to the associated service. For example, an 800 service has an 800-type trigger at the switching system, an 800-service database at the SCP, and an 800-service management system to support the 800 SCP. In this service-specific environment, the 800-service set of capabilities cannot be used for other services (e.g., 900 service). Although the service logic is external to the switching system, it is still service-specific.

At first glance, Figure 1.1.4 looks similar to the previous diagram. However, there is one fundamental difference. Notice the wording "service-independent management systems" under the box labeled "service management system." Now, following the IN/1 800 service-specific example, the AIN service-independent software has a three-digit trigger capability that can be used to provide a range of three-digit services (800, 900, XXX, etc.) as opposed to 800 service-specific logic. Likewise, the SCP service logic and the service management system are service independent, not service specific. AIN is a service-independent network capability!

1.1.5 Benefits of INs

The main benefit of INs is the ability to improve existing services and develop new sources of revenue. To meet these objectives, providers require the ability to:

Introduce New Services Rapidly
IN provides the capability to provision new services or modify existing services throughout the network with physical intervention.

Provide Service Customization
Service providers require the ability to change the service logic rapidly and efficiently. Customers are also demanding control of their own services to meet their individual needs.

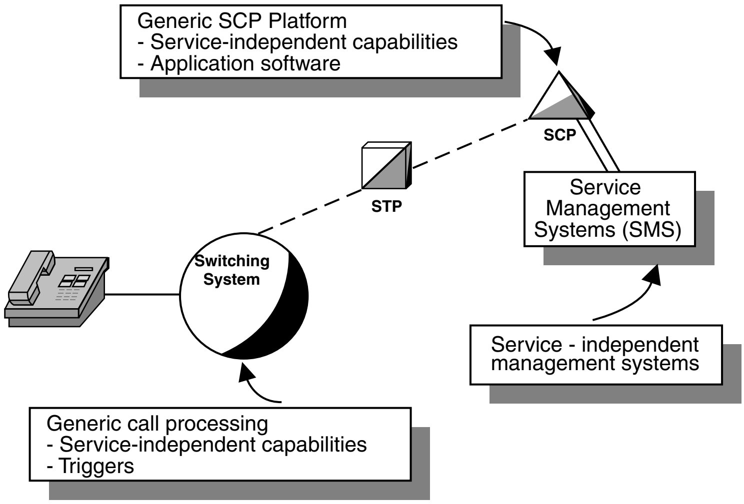

FIGURE 1.1.4 Advanced intelligent network (AIN) architecture.

Establish Vendor Independence
A major criteria for service providers is that the software must be developed quickly and inexpensively. To accomplish this, suppliers have to integrate commercially available software to create the applications required by service providers.

Create Open Interfaces
Open interfaces allow service providers to introduce network elements quickly for individualized customer services. The software must interface with other vendors' products while still maintaining stringent network operations standards. Service providers are no longer relying on one or two vendors to provide equipment and software to meet customer requirements.

AIN technology uses the embedded base of stored program-controlled switching systems and the SS7 network. The AIN technology also allows for the separation of service-specific functions and data from other network resources. This feature reduces the dependency on switching system vendors for software development and delivery schedules. Service providers have more freedom to create and customize services.

The SCP contains programmable service-independent capabilities (or service logic) that are under the control of service providers. The SCP also contains service-specific data that allows service providers and their customers to customize services. With the IN, there is no such thing as one size fits all — services are customized to meet individual needs.

Since service logic is under the service provider's control, it is easier to create services in a cost-effective manner. Network providers can offer market-focused service trials by loading service logic in a SCP and triggering capabilities in one or more switching systems.

Accepted standards and open, well-documented interfaces provide a standard way of communicating between switching systems and SCPs, especially in a multi-vendor environment.

1.1.6 Local Number Portability

The Telecommunications Act of 1996 is having a profound impact on the U.S. telecommunications industry. One area of impact that is being felt by everyone is Local Number Portability (LNP). For LNP, the Federal Communications Commission (FCC) requires the nation's local exchange carriers (LECs) to allow customers to keep their telephone numbers if they switch local carriers. The LECs must continue to maintain the quality of service and network reliability that the customer has always received.

The rules required that all LECs begin a phased deployment of a long-term service provider portability solution no later than October 1, 1997 in the nation's largest metropolitan statistical areas.

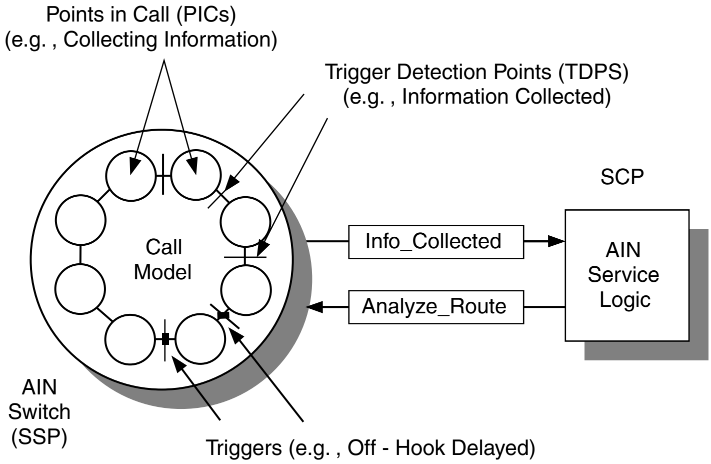

FIGURE 1.1.5 The call model: basic concept.

Wireless carriers are also affected by LNP. December 31, 1998 was the deadline date that wireless carriers had to be able to complete a call to a ported wire–line number. By June 30, 1999, the Act called for full portability between wireless and wireline, including roaming capabilities.

AIN is a logical technology to help service providers meet this mandate. Many providers are looking to AIN LNP solutions because of the flexibility that AIN provides without the burden of costly network additions.

1.1.7 The Call Model

The call model is a generic representation of service switching point (SSP) call processing activities required to establish, maintain, and clear a basic call. The call model consists of Point in Calls (or PICs), Detection Points (DPs), and triggers. These are depicted in Figure 1.1.5.

PICs represent the normal switching system activities or states that a call goes through from origination to termination. For example, the null state or the idle state is when the SSP is actually monitoring the customer's line. Other examples of states, or PICs, are off-hook (or origination attempt), collecting information, analyzing information, routing, alerting, etc.

Switching systems went through similar stages before AIN was developed. However, the advent of AIN introduced a formal call model that all switching systems must adhere to. In this new call model, trigger detection points (TDPs) were added between the PICs. SSPs check TDPs to see if there are any active triggers.

There are three types of triggers: subscribed or line-based triggers, group-based triggers, and office-based triggers. Subscribed triggers are provisioned to the customer's line, so that any calls originating from or terminating to that line would encounter the trigger. Group-based triggers are assigned to groups of subscribers, e.g., business or Centrex groups. Any member of a software-defined group will encounter the trigger. Office-based triggers are available to everyone connected to the telephone switching office or has access to the North American numbering plan. Office-based triggers are not assigned to individuals or groups.

If an active trigger is detected, normal switching system call processing is suspended until the SSP and SCP complete communications. For example, in Figure 1.1.5, suppose an AIN call has progressed through the null state or PIC, the off-hook PIC, and is currently at the collecting information PIC. Normal call processing is suspended at the information collected TDP because of an active off-hook delayed trigger. Before progressing to the next (analyzing information) PIC, the SSP assembles an information collected message and sends it to the SCP over the SS7 network. After SCP service logic acts on the message, the SCP sends an analyze route message that tells the SSP how to handle the call before going to the next PIC (analyzing information).

Essentially, when the SSP recognizes that a call has an associated AIN trigger, the SSP suspends the call processing while querying the SCP for call routing instructions. Once the SCP provides the instruction, the SSP continues the call model flow until completion of the call. This is basically how a call model works, and it is a very important part of AIN.

This concept differs from the pre-AIN switching concept in which calls were processed from origination state to the call termination state without call suspension.

1.1.8 AIN Releases

The demand for AIN services far exceeded the availability of network functionality. Service providers could not wait for all the features and functionality as described in AIN Release 1. AIN Release 1 defined all types of requirements, which made the capability sets too large to be adopted by the industry.

In North America, the industry agreed to develop subsets of AIN Release 1 that provided for a phased evolution to AIN Release 1. AIN 0.1 was the first subset targeted for use.

Bellcore developed functionality to address the FTS 2000 requirements set forth by the U.S. Government. The RBOCs AIN turn adopted these requirements to meet their customers' immediate needs. This effort resulted in AIN Release 0, which had a time frame before the availability of AIN 0.1.

Meanwhile, the global standards body, the International Telecommunications Union (ITU), embraced the concepts put forth in the AIN Release 1 requirements. The ITU developed an international IN standard called Capability Set 1, or CS-1. As with AIN Release 1 in North America, CS-1 was encompassing a rich functionality. To meet the market demand, the ITU formed a subgroup called European Telecommunications Standards Institute (ETSI) to focus on the immediate needs. This subgroup developed the Core INAP capabilities. Many Post Telegraph and Telecommunications (PTT) organizations and their switch vendors have adopted the ETSI Core INAP as the standard and are providing Core Intelligent Network Application Protocol (INAP) capabilities.

1.1.8.1 AIN Release 1 Architecture

Figure 1.1.6 shows the target AIN Release 1 architecture, as defined in Bellcore AIN Generic Requirements (GRs).

The SSP in this diagram is an AIN-capable switching system. In addition to providing end users with access to the network and performing any necessary switching functionality, the SSP allows access to the set of AIN capabilities. The SSP has the ability to detect requests for AIN-based services and establish communications with the AIN service logic located at the SCPs. The SSP is able to communicate with other network systems (e.g., intelligent peripherals) as defined by the individual services. The SCP

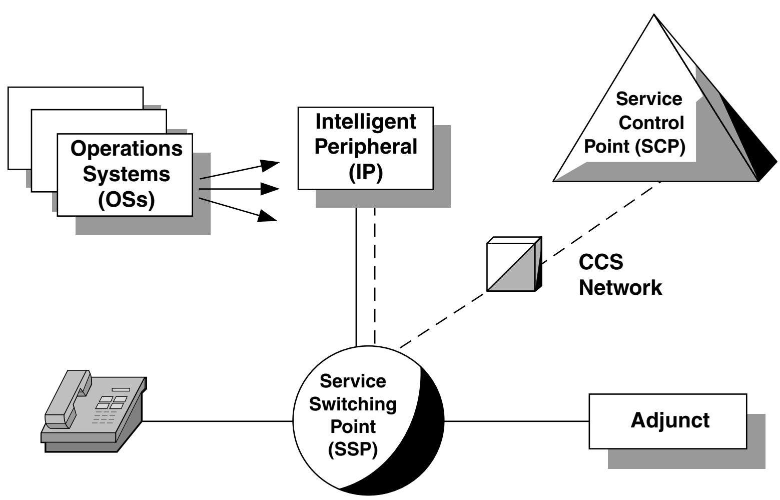

FIGURE 1.1.6 AIN Release 1.

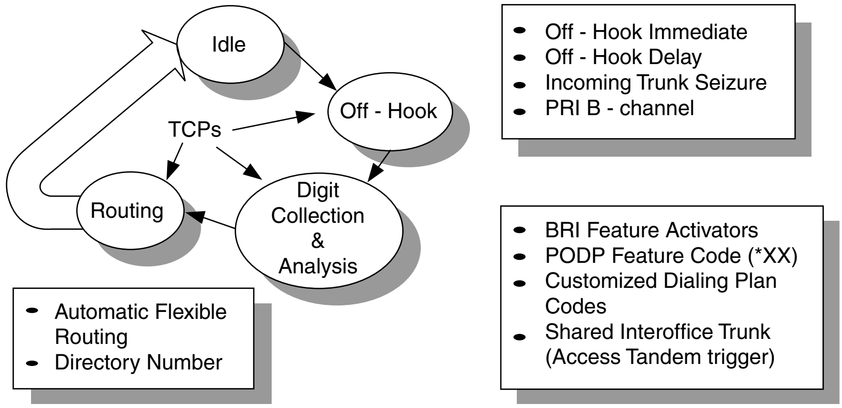

FIGURE 1.1.7 AIN Release 0 call model.

provides the service control. There are two basic parts to a SCP. One part is the application functionality in which the service logic is installed after the services have been created. This application functionality sits on top of the second basic SCP part: a set of generic platform functionalities that are developed by SCP vendors. This platform functionality is shared among the service logic application programs in the application functionality. The platform functionality also provides the SS7 interface to switching systems. As shown in Figure 1.1.6, the SCP is connected to SSPs by the SS7 network.

The intelligent peripheral (IP) provides resources such as customized and concatenated voice announcements, voice recognition, and dual tone multi-frequencies (DTMF) digit collection. The IP contains a switching matrix to connect users to these resources. In addition, the IP supports flexible information interactions between an end user and the network. It has the resource management capabilities to search for idle resources, initiate those resources, and then return them to their idle state.

The interface between the SSP and the IP is an integrated services digital network (ISDN), primary rate interface (PRI) and/or basic rate interface (BRI). The IP has the switching functionality that provides the ISDN interface to the switching system. The adjunct shown in Figure 1.1.6 is functionally equivalent to a SCP, but it is connected directly to a SSP. A high-speed interface supports the communications between an adjunct and a SSP. The application-layer messages are identical in content to those carried by the SS7 network between the SSP and SCP.

1.1.8.2 AIN Release 0

The AIN Release 0 call model has three trigger checkpoints (TCPs). At each TCP there are one or more triggers. For example, the off-hook TCP includes the off-hook immediate trigger. If a subscriber's line is equipped with this trigger, communications with the SCP will occur if the switching system detects an off-hook condition. For an off-hook delayed trigger, one or more digits are dialed before triggering to the SCP. At the digit collection and analysis TCP, collected digits are analyzed before triggering. Triggering may also occur at the routing stage of a call. This call model is shown in Figure 1.1.7.

When a switching system recognizes that a call needs AIN involvement, it checks for overload conditions before communicating with the SCP. This process is called code gapping. Code gapping allows the SCP to notify the switching system to throttle back messages for certain NPAs or NPA-NXXs. When code gapping is in effect, some calls may receive final treatment. For others, a provide instruction message is sent to the SCP. Depending on the SCP service logic, it will respond to the switching system with any of the call processing instructions shown in Figure 1.1.8.

AIN Release 0 provided 75 announcements at the switching system. Release 0 was based on American National Standards Industry (ANSI) Transaction Capability Application Part (TCAP) issue 1. TCAP is at layer 7 of the SS7 protocol stack. This means that there is only one message sent from the SSP to the SCP, no matter what trigger is hit at any of the three TCPs.

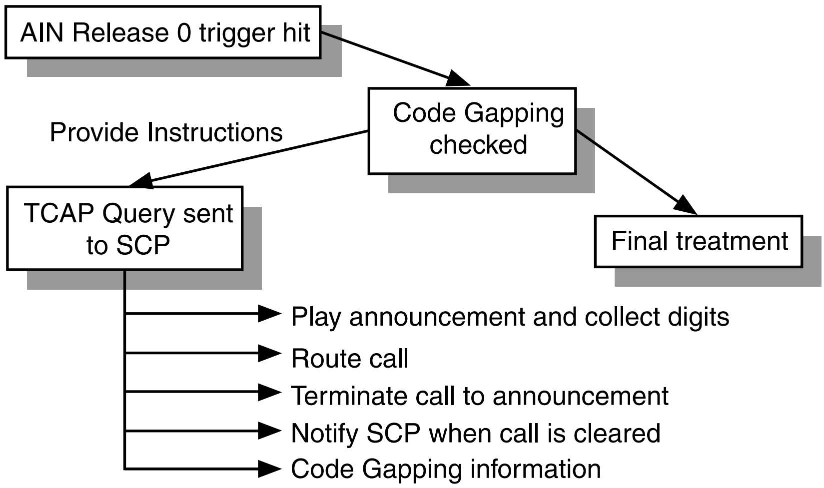

FIGURE 1.1.8 AIN Release 0 functions.

1.1.8.3 AIN Release 0.1

AIN 0.1 is the first subset of AIN Release 1. There are two fundamental differences between AIN Release 0 and AIN 0.1 The first is a formal call model and the second is the messaging sets between the switching system and the SCP. The formal call model is separated into the originating call model (originating half call) and the terminating call model (terminating half call). The AIN Release 0 call model did not distinguish between originating and terminating. A standard or formal call model is necessary as we evolve to the Target AIN Release 1 capability, because the capabilities will have more PICs and TDPs. Also, there will be multiple switch types and network elements involved. Therefore, the service logic will need to interact with every element that will be required in the network.

AIN 0.1 includes several other major features. There are 254 announcements at the switching system, which provide more flexible messages available to customers. There are additional call-related and non-call-related functions as well as three additional triggers — the N11 trigger, the 3–6–10-digit trigger, and the termination attempt trigger. More triggers provide additional opportunities for SCP service logic to influence call processing. (Note: TCP was an AIN Release 0 term that changed to TDP in AIN 0.1). There are several AIN 0.1 non-call-related capabilities. The SCP has the ability to activate and deactivate subscribed triggers. The AIN 0.1 SCP can also monitor resources. In addition to sending a call routing message to the switching system, the SCP may request that the switching system monitor the busy/idle status of a particular line and report changes. AIN 0.1 also supports standard ISDN capabilities.

As mentioned previously, there is a distinction between the originating side and the terminating side of a service switching point. This means that both originating and terminating triggers and service logic could influence a single call. Figure 1.1.9 shows a portion of the AIN 0.1 originating call model. The AIN 0.1 originating call model includes four originating trigger detection points — origination attempt, information collected, information analyzed, and network busy.

The AIN 0.1 terminating call model includes one TDP — termination attempt, as depicted in the partial call model in Figure 1.1.10.

1.1.8.4 AIN 0.1: SSP–SCP Interface

The AIN 0.1, as shown in Figure 1.1.11, is based on ANSI TCAP issue 2, which means that the message set is different than the message set in ANSI TCAP issue 1. For example, in AIN Release 0, there is only one message sent from the SSP to the SCP no matter what trigger is hit at any of the three TCPs. In AIN 0.1, separate messages are sent for the four originating and one terminating TDP.

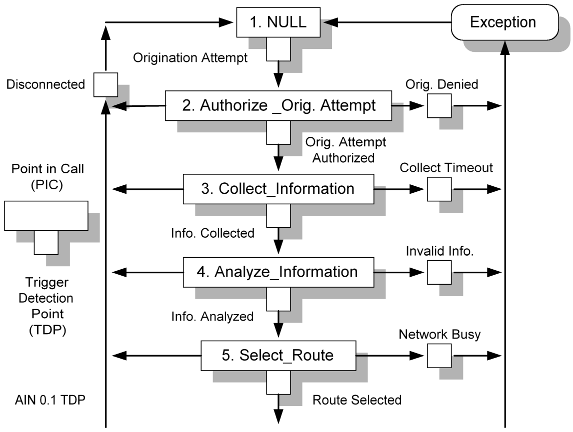

FIGURE 1.1.9 AIN 0.1 originating call model.

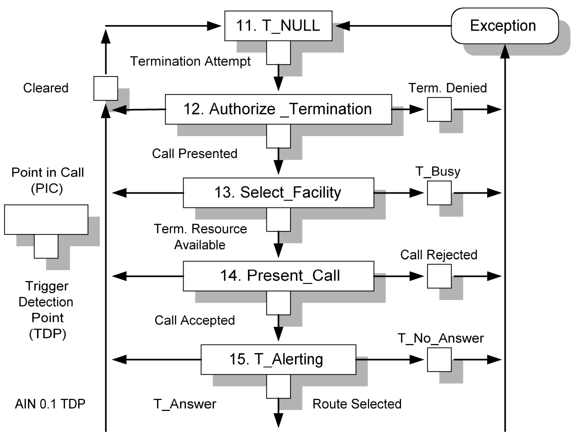

FIGURE 1.1.10 AIN 0.1 terminating call model.

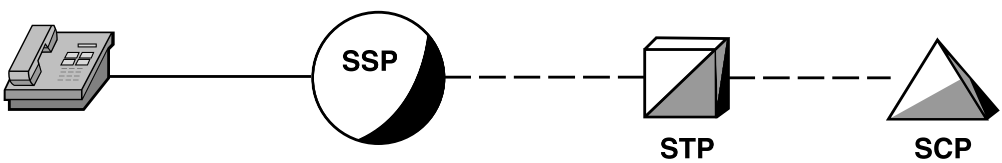

FIGURE 1.1.11 AIN 0.1 SSP–SCP interface.

1.1.8.5 AIN Release 0.2

AIN 0.2 builds on AIN 0.1 with additional capabilities to support two service drivers — Phase 2 personal communication service (PCS) and voice activated dialing (VAD). While AIN 0.2 is focused on capabilities to support PCS and VAD, all requirements for these capabilities are defined in a service-independent manner. AIN 0.2 capabilities will include:

- ISDN-based SSP–IP interface
- Busy and no-answer triggers

- Next event lists processing
- Default routing, and
- Additional functions in all operations areas (e.g., network testing).

The two primary AIN 0.2 capabilities are the ISDN interface between a switching system and an ISDN-capable device (such as an IP) and the addition of busy and no-answer triggers.

Next event lists processing is another important capability. In addition to TDPs, AIN 0.2 includes event detection points (EDPs). With EDPs, the SCP will have the ability to send a next event list to the SSP. This next event list is used by the SSP to notify the SCP of events listed in the next event list. These events may include busy, no answer, terminating resource available, etc.

AIN 0.2 also includes default routing capabilities. This means that when calls encounter error conditions, they can be sent to a directory number, an announcement, etc., as opposed to sending it to final treatment, as is the case in AIN 0.1.

1.1.8.6 AIN 0.2 SSP–IP Interface

AIN Release 0 and AIN 0.1 assumed that the announcements were switch-based. With the introduction of AIN 0.2, announcements can reside in an external database, such as an IP. If the SCP sends a send-to-resource message to the switching system to have the IP play an announcement or collect digits, the switching system connects the customer to the IP via the SSP–IP ISDN interface. The end user exchanges information with the IP. The IP collects the information and sends it to the switching system. The switching system forwards the information to the SCP. One of the fundamental switching system capabilities is the interworking of SS7 (SCP) messages with ISDN messages (SSP–IP).

In addition the SSP may control IP resources without SCP involvement. VAD is an example. A VAD subscriber could be connected to the IP voice recognition capabilities upon going off-hook. The VAD subscriber says "call mom," and the IP returns mom's telephone number to the switching system. The switching system recognizes mom's number as if the subscriber had actually dialed the number.

1.1.9 AIN Service Creation Examples

The previous modules addressed the architecture and the theory of the AIN. This section will discuss various aspects of service creation — the tool that builds the representation of the call flow for each individual customer. Many AIN software vendors have paired service creation software with state-of-the-art computer graphics software to eliminate the need for traditional programming methods. Through the use of menu-driven software, services are created by inputting various service parameters.

1.1.9.1 Building Block Approach

Figure 1.1.12 provides an example of a building-block approach to creating AIN services. Play announcement, collect digits, call routing, and number translation building blocks are shown here. The SSP has the ability to play announcements and collect digits, as does the IP. Routing the call is a SSP function, and number translation is a SCP capability. By arranging these four capabilities or building blocks in various combinations, services such as 800 calling with interactive dialing, outgoing call screening, and area number calling can be created.

1.1.9.2 Service Creation Template

Figure 1.1.13 represents what a service creation template might look like. For an outgoing call screening service, the process begins with the customer's telephone number.

This example allows the customer to screen 900 numbers, while still having the ability to override 900 screening by entering a PIN. Except for 703-974-1234, all non-900 calls are processed without screening.

1.1.9.3 Digit Extension Dialing Service

A 5-digit extension dialing service is displayed in Figure 1.1.14. It allows for abbreviated dialing beyond central office boundaries. If an employee at location 1 wants to call an employee at location 2 by dialing the extension number 1111, 21111 would be dialed.

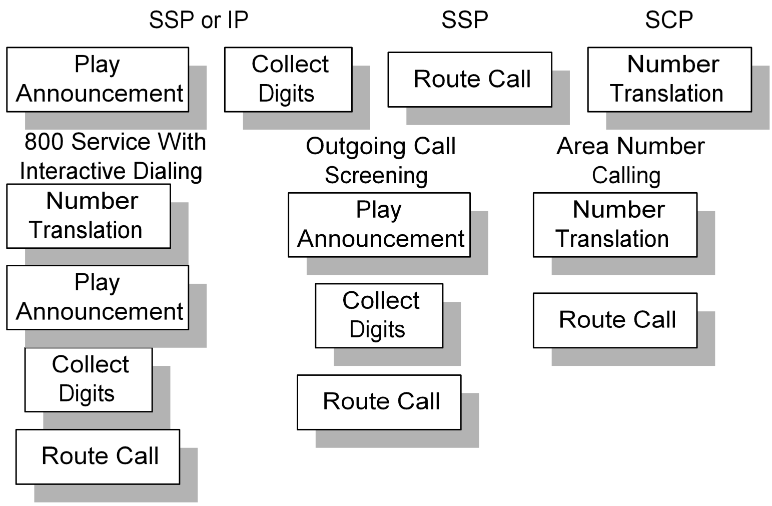

FIGURE 1.1.12 AIN service example: building block approach.

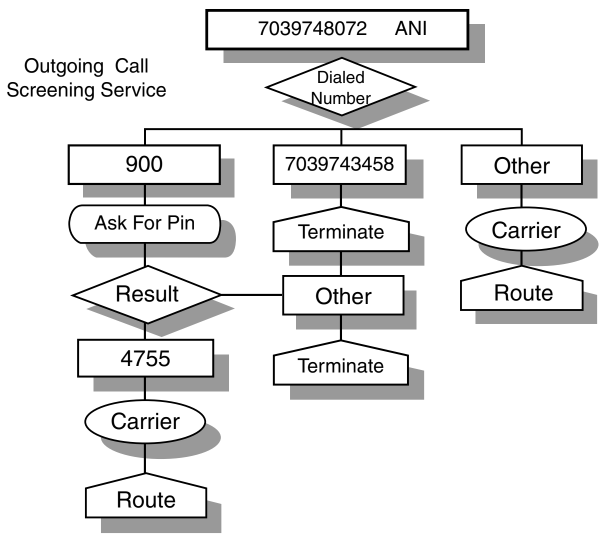

FIGURE 1.1.13 AIN service example: building block approach.

Although 21111 is not a number that a switching system can use to route the call, a customized dialing plan trigger is encountered after 21111 is dialed and a query is sent to the SCP. Service logic at the SCP uses the 21111 number to determine the "real" telephone number of the called party.

1.1.9.4 Disaster Recover Service

Figure 1.1.15 illustrates a disaster recovery service. This service allows businesses to have calls routed to one or more alternate locations based on customer service logic at the SCP. Calls come into the switching system served by the normal location. After triggering, communication with the SCP occurs. Based on the service logic, the call could be either routed to the normal business location or to one or more alternate business locations.

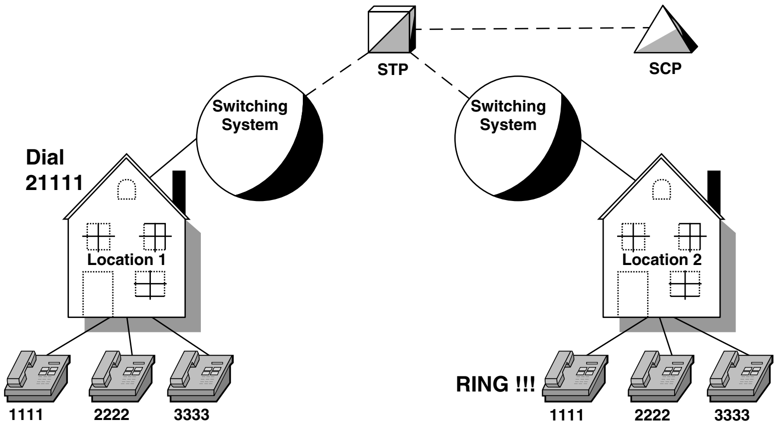

FIGURE 1.1.14 AIN service example: building block approach.

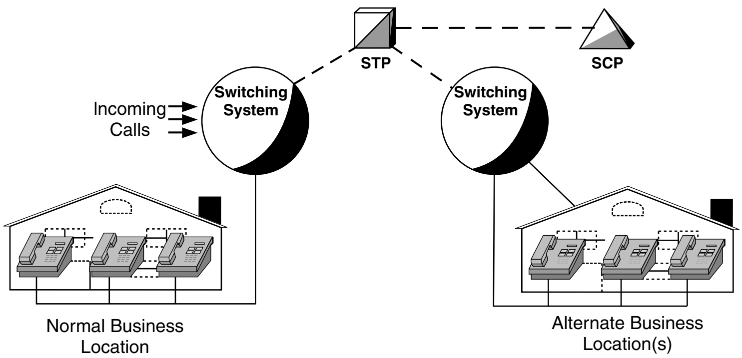

FIGURE 1.1.15 Disaster recovery service.

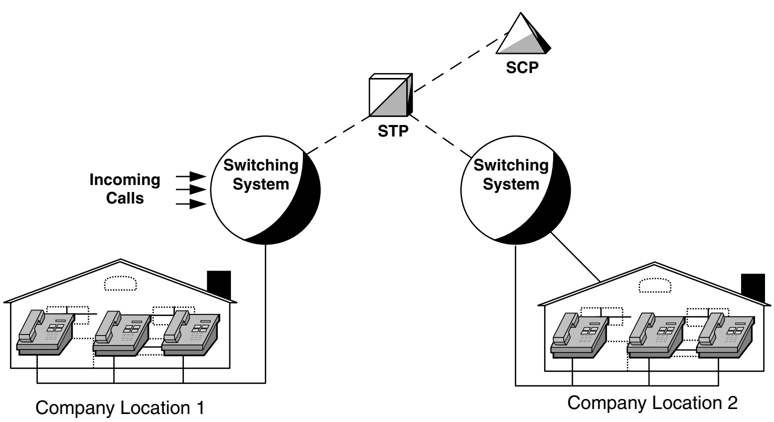

FIGURE 1.1.16 Area number calling (ANC) service.

1.1.9.5 Area Number Calling Service

An area number calling (ANC) service is shown in Figure 1.1.16. This service is useful for companies or businesses that want to advertise one telephone number but want their customer's calls routed to the nearest or most convenient business location. The SCP service logic and data (e.g., zip codes) are used

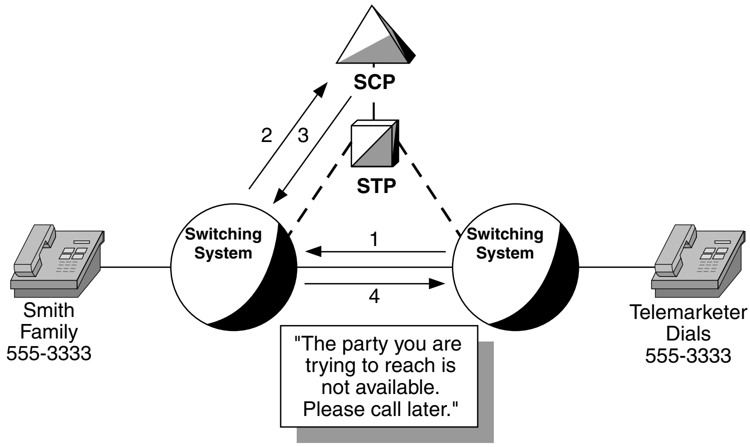

FIGURE 1.1.17 Do not disturb service.

to match the calling party's telephone number and their geographical location. The call is then routed to the company or business location that is closest to or most convenient for the calling party.

1.1.9.6 Do Not Disturb Service

Finally, a do not disturb service is displayed in Figure 1.1.17. This is a service in which the Smith family has terminating screening service logic at the SCP. Whenever someone calls them, the service logic determines whether the call should be routed to the Smith's telephone or play an announcement. In this particular case, a telemarketer calls the Smiths. The SCP tells the switching system to route the telemarketer to an announcement. The customer's SCP service logic may also contain a list of numbers that they want to get through while do not disturb is active. In that case, if the SCP finds a match between the calling party number and a number on the list, the call is routed to the Smiths.

1.1.10 Other AIN Services

The following list describes the services that companies have developed using AIN/IN technology. Some services are tariffed, deployed in the network, and generating revenues. Others are in market or technical trials, getting ready for deployment. There are other services that are either planned for deployment or were developed for demonstration purposes.

N11 access service: With this service, a unique code is used to access a service gateway to information service providers (ISPs), such as newspapers or libraries. The subscriber may either preselect an ISP for automatic routing or request block calls to ISPs.

Basic routing: Allows the subscriber to route calls to a single destination as defined in the system.

Single number service: Allows calls to have different call treatments based on the originating geographical area and the calling party identification.

Routing by day of week: Allows the service subscriber to apply variable call routings based on the day of the week that the call is placed.

Routing by time of day: Allows service subscribers to apply variable call routings based on the time of the day that the call is made.

Selective routing: Tied to the call forwarding feature generally offered as a switch-based feature. With the AIN, when a call to a selective routing customer is forwarded, the SCP determines where to route the forwarded call based on the caller's number.

Call allocator: Allows the service subscriber to specify the percentage of calls to be distributed randomly for up to five alternate call handling treatments.

Alternate destination on busy day: Allows the service subscriber to specify a sequence of destinations to which calls will be routed if the first destination is busy.

Command routing: A service subscriber predefines a set of alternate call treatments to handle traffic in cases of emergency, unanticipated or anticipated demand peaks, or for any other reason that warrants an alternate call treatment.

Call gate: This is a versatile outgoing call screening service. Call gate supports a personal identification number (PIN) and screening based on time of day and day of week.

Personal access: A type of "follow me" service. A virtual telephone number is assigned to the personal access service subscriber. When a caller dials this number, the software determines how to route the call.

Calling party pays: A service offered to cellular customers. It notifies the calling party that they are trying to reach a cellular number. If they choose to complete the call, they will incur the connect charge of the called party. If they elect not to incur the cost, the call may either be terminated or routed to the called party's voice mail.

Remote access to call forwarding (ultraforward): Allows remote access to call forwarding. Callers may, from any location in the world, call in remotely and activate and/or change their call forwarding number.

Portable number service: PNS Features enhanced call forwarding for large business subscribers. It provides subscribers with the ability to maintain a personal itinerary which includes time-of-day, day-of-week (TOD/DOW) schedules, call searching schedules, and call routing information. PNS subscribers also have the ability to override their schedules with default routing instructions. This service is intended for companies with employees who are in highly mobile environments, requiring immediate availability.

Enhanced 800 service: (Freephone) A customer's call to an 800 service subscriber can be routed to different destinations; instances of routing include the geographical location of the caller, the time and day the call is made, and the caller responses to prompts. The subscriber sets alternate routing parameters for the call if the destination is busy or unavailable, thereby redirecting and allowing for completion of the call.

Mass calling service: MCS A polling and information service that permits simultaneous calling by a large number of callers to one or more telephone numbers. MCS provides a variety of announcement-related services that connect a large number of callers (who dial an advertised number) to recorded announcement devices. Two types of offerings are mass announcements, such as time and weather, and televoting, which allows callers to register their opinions on a topic of general interest.

Automatic Route Selection/Least Cost Routing: Subscribers design a priority route for every telephone number dialed. The system either directs calls or blocks calls to restricted privilege users.

Work-at-home: Allows an individual to be reached at home by dialing an office number, as well as allowing the employee to dial an access code from home, make long distance calls, and have them billed and tracked to a business telephone number.

Inmate service: Routes prisoners' calls, tracks the call information, and offers call control features such as prompts for PINs, blocking certain called numbers, and time or day restrictions.

Holding room: Transportation companies' passengers use this service to inform families or business associates of transportation delays or cancellations.

Call prompter: Allows a service subscriber to provide an announcement that requests the caller to enter a digit or series of digits via a dual tone multi-frequency (DTMF) telephone. These digits provide information used for direct routing or as a security check during call processing.

Call counter: Increases a counter in the televoting (TV) counting application when a call is made to a televised number. The counts are managed in the SCP, which can accumulate and send the results during a specific time period.

500 access service: Allows personal communications service (PCS) providers the ability to route calls to subscribers who use a virtual 500 number.

PBX extend service: Provides a simple way for users to gain access to the Internet network.

Advertising effectiveness service: Collects information on incoming calls (for example, ANI, time, and date). This information is useful to advertisers to determine customer demographics.

Virtual foreign exchange service: Uses the public switched network to provide the same service as wired foreign exchange service.

ACNA originating line blocking: ACNA (Automated Customer Name and Address), with the ability to block their line from being accessed by the service.

AIN for the case teams: Allows technicians to dial from a customer premise location anywhere in the service region and connect to a service representative supported by an ACD. Through voice prompts, the technician is guided to the specific representative within a case team pool within seconds, with no toll charges to the customer.

Regional intercept: Instructs callers of new telephone numbers and locations of regional customers. This service also forwards calls to the new telephone number of the subscriber. Various levels of the service can be offered, based upon the customer's selection.

Work at home billing: A person who is working at home dials a 4-digit feature access code which prompts the system to track and record the billing information for the calls. Calls tracked in this manner are billed directly to the company rather than to the individual.

Inbound call restriction: Allows a customer to restrict certain calls from coming into the subscriber's location. This service is flexible enough to restrict calls either by area code, NNX, or particular telephone numbers. Restrictions may even be specified by day of week or time of day.

Outbound call restriction: Allows a customer to restrict certain calls from being completed from the subscriber's location. This service is flexible enough to restrict calls by either area code, NNX, or particular telephone numbers. Restrictions may even be specific to day of week or time of day.

Flexible hot line: Allows a customer to pick up a telephone handset and automatically connect to a merchant without dialing any digits. An example of this is a rent-a-car phone in an airport, which allows a customer to notify the rent-a-car company to pick them up at the terminal.

Acronyms

ABS	Alternative billing source
AIN	Advanced intelligent network
AMP	AIN maintenance parameter
API	Applications programming interface
ASE	Application service elements
BCSM	Basic call state model
BRI	Basic rate interface
BSTP	Broadband signaling transfer point
CCM	Call control module
CCSN	Common channel signaling network
CFM	Call failure message
CSM	Call segment model
DAA	Directory assistance automation
DAP	Data access point
DCN	Data communications network
DP	Detection point
DTMF	Dual tone multi-frequencies
EDP	Event detection point
EML	Element management layer
ETC	Event trapping capability
FE	Functional entity
GDI	Generic data interface
IF	Information flow
IN	Intelligent network
INAP	Intelligent network application protocol

IN/1	Intelligent Network 1
IP	Intelligent peripheral
IPC	Intelligent peripheral controller
IPI	Intelligent peripheral interface
ISCP	Integrated service control point
LEC	Local exchange carriers
LIDB	Line information database
LNP	Local number portability
MP	Mediation point
MSC	Message sequence chart
NAP	Network access point
NCAS	Non-call associated signaling
NCP	Network control point
NE	Network element
NEL	Next event list
NML	Network management layer
NNI	Network-to-network interface
OBCM	Originating basic call model
ONA	Open network architecture
OOP	Object-oriented programming
OPC	Originating point code
PCS	Personal communications service
PIC	Points in call
PP	Physical plane
RBOC	Regional bell operating companies
RDC	Routing determination check
RE	Resource element
RVT	Routing verification test
SCE	Service creation environment
SCMS	Service creation and maintenance system
SCP	Service control point
SDP	Service data point
SIBB	Service-independent building block
SLE	Service logic editor
SLEE	Service logic execution environment
SLI	Service logic interpreter
SLL	Service logic language
SLP	Service logic program
SM	Session management
SMS	Service management system
SN	Service node
SOP	Service order provisioning
SP	Service plane
SPC	Stored program control
SSP	Service switching point
STP	Signalling transfer point
TBCM	Terminating basic call model
TCP	Trigger check point
TCP	Test call parameter
TDP	Trigger detection point
TSC	Trigger status capability
WIN	Wireless intelligent network

References

1. *http://www.telecordia.com*
2. Uyless D. Black, *The Intelligent Network: Customizing Telecommunication Networks and Services,* Prentice Hall Series in Advanced Communications Technologies, 1998.

Further Readings

1. Uyless D. Black, *The Intelligent Network: Customizing Telecommunication Networks and Services*, Prentice Hall Series in Advanced Communications Technologies, 1998.
2. Bill Douskalis, *IP Telephony*, Hewlett Packard Professional Books, Prentice Hall PTR, 2000.
3. William Stallings, *High Speed Networks: TCP/IP and ATM Design Principles*, Prentice Hall, 1997.
4. Kornel Terplan, *Telecom Operations Management Solutions with NetExpert*, CRC Press, 1998.
5. Uyless Black, *ISDN and SS7: Architectures for Digital Signaling Networks*, Prentice Hall, 1997.
6. John G. van Bosse, *Signaling in Telecommunication Networks*, Wiley & Sons, 1997.
7. Paul Ferguson and Goeff Huston, *Quality of Service, Delivering QoS on the Internet and Corporate Networks*, Wiley & Sons, 1998.
8. Daniel Minoli and Emma Minoli, *Delivering Voice over IP Networks*, Wiley & Sons, 1998.

1.2 Computer Telephone Integrated (CTI)

Michel Gilbert

1.2.1 Abstract

In the universe of telecommunications, the worlds of voice and data have long been resistant to unification. The basic principles that underlie the two worlds have led to, at best, an uneasy truce. In recent times, however, integration has become the buzzword. The industry has seen the emergence of one technology after another that attempts to draw these two domains into closer proximity. Computer–telephone integration (CTI) is yet another arena in which data and voice encounter one another. In the CTI arena, however, voice and data appear to be on the cusp of a working relationship. This paper introduces and reviews the concepts that underlie the world of CTI, the elements that comprise a CTI application, and the standards that have emerged.

1.2.2 Basic Definitions

In a 1990 article titled "PBX/Host Interfaces: What's Real, What's Next" (Probe Research Conference Digest), Lois B. Levick of Digital Equipment Corporation defined CTI as, "A technology platform that merges voice and data services at the functional level to add tangible benefits to business applications." There are four key elements to this definition: 1) identifying CTI as a technology, 2) a focus on the integration of voice and data, 3) specifying a functional integration, and 4) the need to derive tangible benefits in a business environment.

First, some would dispute the notion that CTI is a new technology. They would suggest that CTI is actually a new application for pre-existing technologies. This is indeed the case. Not only is CTI simply a place to reuse existing technologies, it is also not (as we shall see) particularly new.

Second, the integration of voice and data is a key element in CTI, as the name itself implies. CTI builds on some remarkable convergence points in the evolution of computing and telephony. One of the earliest telephone exchanges was designed in 1889 by a frustrated funeral director! Almond B. Strowger was tired of seeing his competitor get the bulk of the funeral business by virtue of the fact that his competitor's spouse happened to operate the local telephone exchange. To deal with the problem, Strowger designed a telephone exchange that became generally known as a step-by-step (or stepper) exchange. Fifty-four years later, with funding from IBM, Howard Aiken created the Harvard Mark I. Both systems were entirely electromechanical, monstrous in size, and highly rigid in their design. Over the years, however, both computers and switches became entirely electronic and based on solid-state technologies.

Where early switches and computers tended to be hardwired, modern switches and computers are both stored-program machines and very flexible. The switch uses a stored-program model to handle call routing operations. The computer uses a variety of stored programs to support end-user applications. Both depend on a data communications infrastructure to exchange control information. Finally, the telephone network is rapidly converging to the digital communications model, which computers have used almost from the outset.

Telephone switches have become specialized computers designed to provide a switching function, and exchanging information via a complex digital data communications infrastructure.

The third major part of the definition, functional integration, requires a brief sidetrack to examine the anatomy of a phone call. A phone call can be divided into two logical activities, commonly referred to as call control and media processing. Call control is concerned with originating, maintaining, and terminating a call. It includes activities like going off-hook, dialing the phone, routing a call through a network, and terminating a call. Media processing is concerned with the purpose of the phone call. It deals with the type of information being conveyed across the call, and the format in which that information is presented.

Functional integration means the computer and switch collaborate in call control and/or media processing operations. They may actually interchange functions to meet the needs of an application. Data stored in the computer might be useful for routing incoming and/or outgoing calls. Perhaps the simplest example is an autocall application where the user can click on a name stored in a local application and the computer retrieves the associated phone number and dials the call automatically. Alternatively, call-related data can be used to trigger information retrieval from the computer. For example, automatic number identification (ANI) can provide the calling number, which can be used to key a database lookup to retrieve a particular customer's account information before the phone even rings. In both examples, the data of the computer and the routing of a call are bound together to do work.

Another form of functional integration is when computer and telephone peripherals begin to be used interchangeably. For example, computer peripherals can become alternative call control elements instrumental in call monitoring, and telephone network peripherals can become an alternative method for moving data between people and computers. There is even a degree of functional integration achieved when the computer and telephone system are managed from a single point.

The fourth and final element of Levick's definition concerns the benefits CTI brings to business applications. One of the obvious goals of any business application is to provide better service to customers. CTI can increase responsiveness, reduce on-hold waiting times, provide the customer with a single point of contact, and make it easier to provide a broader range of services.

CTI can also increase effectiveness by eliminating many of the mechanical tasks associated with telephony (e.g., dialing phones, looking up phone numbers, etc.), providing a better interface to the telephone system, and integrating control of the phone system into a familiar and regularly used computer interface (e.g., the familiar Windows desktop).

Perhaps the most telling benefit CTI brings to the corporate world (and the one most likely to garner the attention of the decision makers) is the potential for reductions in operating costs. Correctly applied, CTI can mean faster call handling, which translates to reduced call charges. Automation of call-related tasks means potentially fewer personnel, or greater capacity for business with existing personnel. Some CTI implementers have claimed 30% improvement in productivity.

1.2.3 A Brief History of CTI

Although CTI appears to be a recent introduction into the telecommunications arena, there were attempts to integrate voice and data into competitive business applications as early as the 1960s. In his book *Computer Telephone Integration* (ISBN 0-89006-660-4), Rob Walters describes an application put together by IBM for a German bookstore chain.

The bookstores were looking for a way to automate their ordering process. IBM produced a small, hand-held unit that each store manager could use to record the ISBN numbers of books they needed, together with the desired quantity of each. These small units were then left attached to the telephone at the end of the day. Overnight, an IBM 360 located at company headquarters would instruct the IBM 2570 PABX to dial each store in turn.

Once the connection was formed, the IBM mainframe would download the order and then instruct the PABX to release the connection and proceed to the next store. The link between the IBM 360 and the 2750 PABX was called teleprocessing line handling (TPLH). By the end of the night, the 360 would produce a set of shipping specifications for each store, the trucks would be loaded, and the books delivered.

In 1970, a Swedish manufacturer of ball bearings (SKF) replaced its data collection infrastructure with a CTI application that was also based on the IBM 360/2570 complex. Rather than using data collectors who would travel from shop to shop, local shop personnel provided the data directly. On a daily basis, they would dial a number that accessed the IBM 360/2750 complex at headquarters. Data was entered using push-button phones. The switch would pass an indicator of the numbers pressed to the 360 via the TPLH connection, and the computer would return an indication of acceptance or rejection of the data to the switch. The switch would, in turn, produce appropriate tones to notify the user of the status of the information exchange.

These two examples underscore the flexibility of this early system. Note that both outbound (IBM 360 initiates the calls) and inbound (users call the IBM 360) applications were supported. This system exhibited two classic hallmarks of a CTI application. First, the phone connection is used for media processing (i.e., the information being passed back and forth). Second, there is a linkage between the computer and the switch to exert call control.

Amazingly, after IBM's introduction of the 360/2570 applications, there was an attempt at a form of electromechanical CTI, albeit a short-lived one. In 1975, and largely in response to the IBM 360/2570 solution, the Plessey company designed a computer link to their crossbar PABX. Every line and every control register of the switch was wired to the computer so its status could be monitored and controlled. The computer could intercept dialed digits, make routing decisions, and instruct the switch to route a call in a particular fashion. Called the System 2150, only two were deployed before electronic switching rendered the technology obsolete.

At about the same time, a group of Bellcore researchers formed the Delphi Corporation to build a system for telephone answering bureaus. These bureaus were essentially answering services for multiple companies. At the end of the day, the company phones were essentially forwarded to these bureaus, where a person would answer the line and take a message. However, it was important for the person answering the phone to know what company was being called, and to be able to answer the phone as a representative of that company. Delphi 1, released in 1978, was the answer to the problem.

All calls were rerouted to a computer that could tell by the specific line being rung which company was being called. The computer would then retrieve the text for that company's standard greeting, as well as any special instructions for handling the call, and pass the call and instructions to an attendant. The answering bureaus saw a 30% increase in efficiency and the concept caught on quickly.

Through the 1980s, niche applications continued to appear, and new players entered the market. These included British Telecom (a telemarketing application), Aircall (paging), and the Telephone Broadcasting Systems (a predictive dialing system). Perhaps one of the best-known CTI applications to emerge in the 1980s was Storefinder™. The results of a collaboration between Domino's Pizza and AT&T, Storefinder™ used ANI to route a call to the Domino's Pizza nearest that customer. Before the phone in the store could ring, Storefinder™ provided the personnel at that store with the customer's order history, significantly enhancing the level of customer service.

Many early attempts to integrate computers and telephony focused on the media processing aspect of communication. This includes early versions of voice mail and interactive voice response (IVR) systems. These simple technologies did not need much more than specialized call receiving hardware in a computer system, and a hunt group. When a caller dialed in to the service, the telephone network switched the call to one of the access lines in the hunt group. The computer then proceeded to provide voice prompts to guide the user through the service. In the case of voice mail, the user was prompted to leave or retrieve recorded messages. In the case of IVR, the user was prompted to provide, by touch-tone or voice, the information necessary to perform a database lookup (e.g., current credit card balances, history of charges, mailing address, payment due dates, etc.).

Modern voice-mail and IVR systems, and more advanced CTI applications, include a strong call control component. They can transfer calls, provide outward dialing, and even paging. This requires a more complex physical and logical integration of the computer and telephony worlds. The two worlds must be physically connected, making it possible for data from the telephone network to be passed to the computer and call control information from the computer to be passed to the network. Logically, the integration of data from both the telephone network and the computer must be used to create new applications that give the corporation a competitive edge.

Today, the call center scenario dominates that CTI world. Resulting applications typically utilize the most advanced call control and media processing functions. CTI enables new call center models. A single call center can be logically partitioned to function as multiple smaller call centers, or multiple distributed call centers can be logically integrated to act as one. Modern CTI applications provide the knife, or the glue, to make these models possible.

1.2.4 Components and Models

The basic components of a CTI application are depicted in Figure 1.2.1. At the heart of the application lies the computer and the switch. The computer houses end-user data and hosts the end-user interface to the CTI application. The switch provides the ability to make and receive calls and hosts the network interface to the CTI application. The computer provides a set of peripherals (e.g., keyboard, screen, etc.) by which the user accesses the CTI application, and the switch provides the peripheral (e.g., telephone) by which the user communicates. Between the computer and switch there must exist a connection or link, the nature of which differs depending on the type of CTI application.

Consider the automated attendant application. A person needing to speak with someone within the company dials the company's published phone number. The switch routes the call to a computer that begins to play back a recorded message. The message prompts the caller to use the touch-tone buttons to select from an array of options. The caller can enter the extension of the person they wish to reach, in which case the computer directs the switch to reroute the call to that extension. The caller can use the keypad to enter the name of the person being reached. The computer has to translate each tone to the associated letter values, and determine if there is a match in the company personnel listing. If there is none, or if the match is ambiguous (e.g., "Sam" and "Pam" use the same key combination), the computer asks the caller to hold and transfers the call to an operator. If a single, unambiguous match is found, the

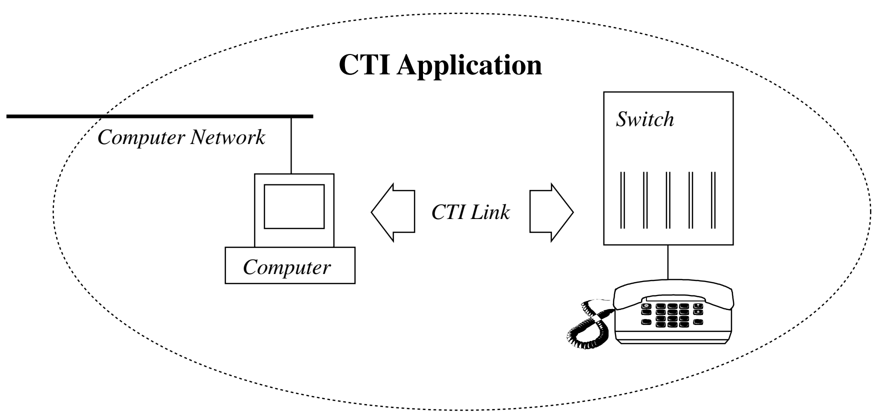

FIGURE 1.2.1 Basic components of a CTI application.

computer can ask the caller to confirm the match, retrieve the extension from the database, and direct the switch to transfer the call. At any point the caller can force the computer to transfer the call to an operator by pressing 0.

1.2.4.1 Media Processing

As has been noted, any phone call can be broken down into two broad activities: media processing and call control. CTI applications typically support both, albeit in different degrees of complexity and by using different strategies. However, a complete suite of CTI services requires both media processing and call control services.

Media processing is perhaps the easiest to understand. When a fax machine calls another fax machine, the transmission of the encoded image across the connection is media processing. When an end user uses their modem to dial in to the local Internet Service Provider (ISP), the exchange of data across the connection is also media processing.

In the CTI arena, the hardware required for media processing is relatively simple. It often takes the form of voice processing, speech digitization and playback, and fax circuitry. Many products integrate these functions into a single printed circuit board that can be installed in a desktop computer. Many of these integrated boards support multiple lines and hardwire the circuitry to each channel. This is sometimes referred to as dedicated media processing hardware (see Figure 1.2.2). Companies that provide such integrated boards include Dialogic Corporation (www.dialogic.com), Pika Technologies, Inc. (www.pika.ca), and Rhetorex (www.rhetorex.com). Rhetorex is now a subsidiary of Lucent Technologies (www.lucent.com).

This approach is appropriate for small-scale applications. For example, a company providing voice mail services in a small town might equip a standard desktop system with a four-line integrated board. A user dialing into the service would be switched by the network to one of the four lines. Based on the

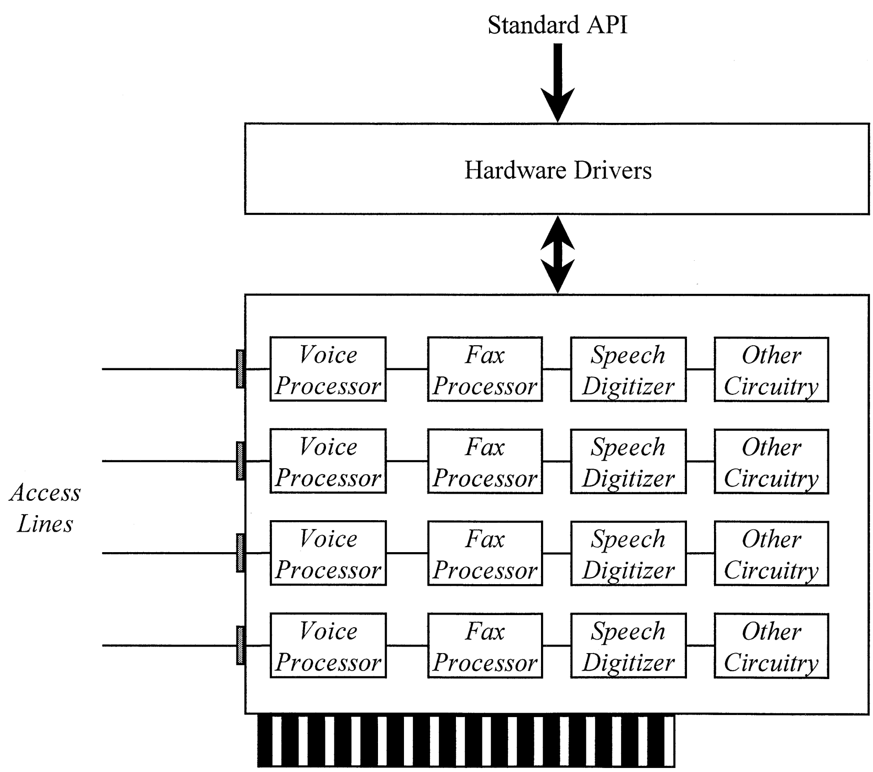

FIGURE 1.2.2 Dedicated media processing hardware.

tones provided by the user (e.g., "Please enter your mailbox number") or ANI information provided by the network, the user can retrieve recorded messages from the computer and play them back.

In these simple environments, standard application programming interfaces (API) are often adequate for controlling the resources. For example, the Microsoft Windows or Solaris APIs that are used to play sound files through a local speaker can also be used to send and receive multimedia content over a telephone connection.

Large-scale applications, however, are more complex. In these environments, sharing resources is more economically viable. A business person may be willing to purchase four complete sets of media processing circuitry, knowing that at any given time only a few components associated with any particular line are going to be used. However, equipping every line in a large application with all of the circuitry it might be called upon to use is not cost effective. For example, consider a large-scale application that implements a pool of four T1 circuit interfaces (96 voice channels). Usage patterns may show that this application needs 96 voice digitizers and playback units, but only 16 speech recognizers, 16 fax processing circuits, and 36 analog interfaces for headsets.

Assembling components at a more modular level is more cost effective and can scale more easily, but it also places new demands on the system. New APIs and standards are required for interconnecting, using, and managing these resources. There are two leading architectures for building such systems: the multi-vendor integration protocol (MVIP) and SCbus. In addition to describing the hardware architecture needed to interconnect telephony-related components, both GO-MVIP and SCSA define software APIs required to use and manage those resources (see Figure 1.2.3). The SCSA Telephony Application Objects (TAO) Framework™ is the API defined by the SCSA.

On the hardware side, both MVIP and SCbus describe a time-division bus for talk-path interconnection, and a separate communication mechanism for coordinating the subsystems. MVIP (www.mvip.org) is administered by the Global Organization for the MVIP (GO-MVIP). SCbus was originally developed by the Signal Computing System Architecture (SCSA™) working group (www.scsa.org). SCSA has since

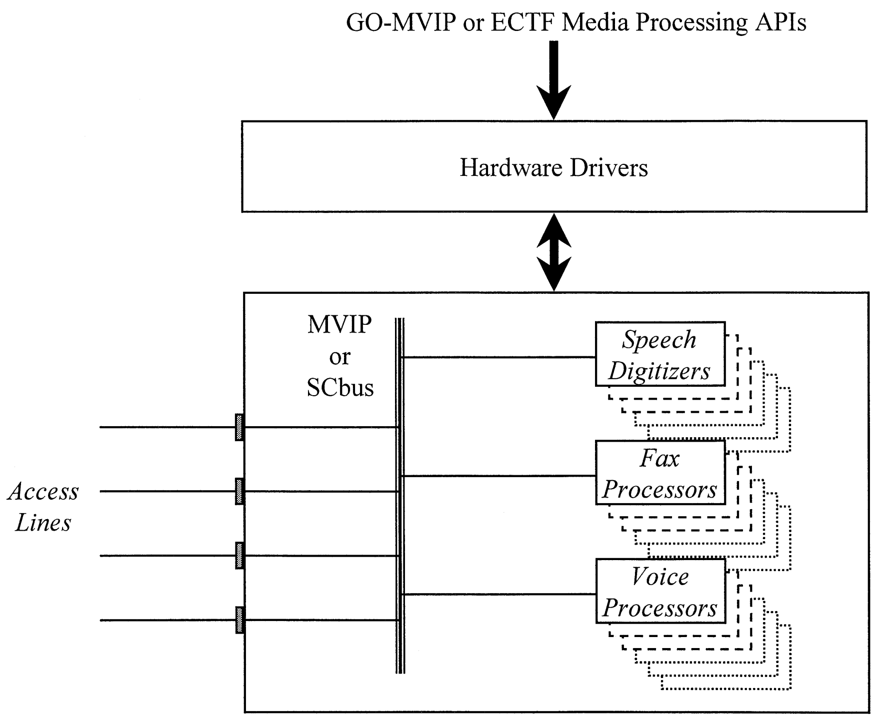

FIGURE 1.2.3 Architecture for sharing media processing hardware.

become part of the Enterprise Computer Telephony Forum (ECTF), a non-profit organization actively prompting the development of interoperability agreements for CTI applications (www.ectf.org). SCbus, announced in 1993, is now also an ANSI standard.

Both GO-MVIP and the ECTF also define a set of application program interfaces (API) for media processing.

1.2.4.2 Call Control

The other major activity a CTI application needs to support is call control. Call control is concerned with the successful establishment, maintenance, and termination of calls. To support these activities, the switching nodes in the telephone network must communicate with one another and with the end-user's terminal equipment. The process by which the switches do this is called signaling. Signaling can be done in-band or out-of-band. In-band signaling occurs on the same channel occupied by user information. This is common for terminal equipment (i.e., telephones), and has become less common within the network itself. Out-of-band signaling occurs on a separate channel from that occupied by user data. This approach is common within the telephone network, and less common between the user and the network (ISDN notwithstanding).

In addition to differentiating between in-band and out-of-band signaling, it is important to note that signaling between the network and the user is bidirectional. The user signals the network by going off-hook, dialing a phone number, and hanging up a phone. This signaling is well standardized. The most common standard today is dual tone multi-frequency (DTMF), the familiar tones we hear as we press buttons on a touch-tone phone. The network signals the user in-band by providing dial tone, busy signals, ringing tones, fast busy, and so forth. Each of these has a distinct meaning, but the sounds have not been well standardized internationally. This is a significant challenge for the CTI environment. Out-of-band network-to-user signaling is somewhat more standardized. Examples include the D-channel on an integrated services digital network (ISDN) interface, the proprietary interfaces defined by digital telephones, and dedicated CTI interfaces to private branch exchanges (PBX) and switches.

Perhaps the most challenging aspect of CTI applications is achieving accurate and reliable call control. In most applications, out-of-band signaling is preferred. Each option, however, has its scope, strengths, and weaknesses. In an ISDN environment, D-channel signaling can be used by the CTI application. One possible CTI application is a network-based automatic call distributor (ACD). Naturally the scope is limited to the domain for which the ISDN signaling is meaningful. For example, the ACD application may not be completely effective when calls cross some public network boundaries.

A CTI application could also leverage the proprietary signaling between a PBX and a digital telephone. Again, such an application may be limited to the scope of the PBX or a group of PBXs from the same manufacturer.

In the public network, the switch-to-switch signaling protocol is called Signaling System 7 (SS7). The domain for SS7 signaling can be as large as an entire public telephone network. Unfortunately, SS7 is usually not available to the CTI application. Closely associated with the internal operation of the public network, SS7 access is jealously guarded by most carriers. Where access is available to the corporate customer, a CTI application based on SS7 requires sophisticated customer premises equipment (CPE) that can handle the complexity of SS7. As a result, this signaling option is usually only appropriate for call centers handling large volumes of calls.

One of the most popular strategies for CTI applications is the dedicated CTI link implemented by many modern PBXs and some public exchange switches. The domain for a dedicated CTI link is a single telephone switch or a small number of tightly integrated switches or PBXs. These facilities are designed for CTI, and tend to offer the rage of signaling options best suited to this environment. These dedicated facilities can implement proprietary or standard call control strategies. Examples of proprietary strategies include Nortel's Meridian Link Protocol (MLP) and AT&T's ASAI Protocol.

Naturally, the industry is leaning strongly to standards-based strategies. The predominant standard is the Computer-Supported Telephony Application (CSTA) from the ECMA (formerly European Computer Manufacturers Association). Adopted in 1990, the CSTA protocol (www.ecma.ch) has now been implemented by

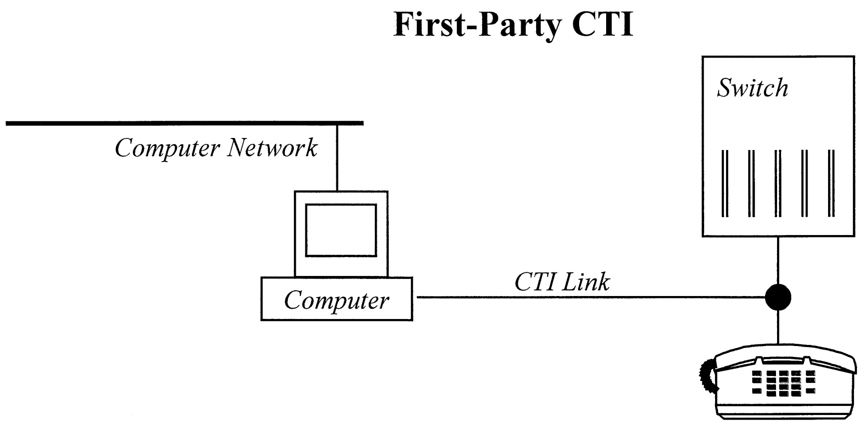

FIGURE 1.2.4 First-party CTI model.

such major players as Siemens ROLM, Ericsson, and Alcatel, to name a few. It is important to note that, although CSTA is a standard, the features any particular vendor elects to implement can vary. As a result, CSTA implementations from different vendors are not necessarily interoperable.

1.2.4.3 First-Party and Third-Party CTI

CTI applications can be broken into two broad classes based on the relationship between the computer and the switch. In first-party CTI, the computer is essentially on an extension to the line on which a call is being received. The computer can exert the same call control functions a human attendant could exert via a standard telephone set attached to the telephone system. This implies that call control is on a call-by-call basis. First-party CTI call control includes such activities as going off-hook, detecting dial tone, dialing a call, monitoring call status signals (e.g., ring, ring no-answer, answer, busy, and fast busy) conditions, and terminating the call.

In the first-party CTI model (Figure 1.2.4) the computer, the keyboard and screen, and the phone are all on the same line. The computer will tend to use the dedicated media processing hardware model, and tend to be a user end-system (as opposed to being a server). First-party CTI is further subdivided into basic and enhanced flavors. Essentially, basic systems use in-band signaling and have limited capability. Enhanced systems use out-of-band signaling, usually either ISDN or proprietary signaling to the PBX. While there are basic first-party CTI platforms on the market, the industry is more interested in enhanced first-party CTI systems.

The classic example of an inbound first-party CTI application is the voice mail system. In a voice mail application, an inbound call is received by the computer. The computer activates the local voice mail software to record and store, or retrieve and playback, voice mail. The simplest example of an outbound first-party CTI application is autocall.

APIs for first-party call control first appeared from the manufacturers of network access equipment (e.g., modems, fax boards, etc.). The only such API that achieved de facto standards status was the Hayes modem command set. Now universally understood by modem products, the Hayes command set defines basic commands for initiating and terminating calls, and altering the configuration of the modem.

Third-party CTI is the more sophisticated model. In third-party CTI, the computer exerts call control via a dedicated connection to the switch or PBX (Figure 1.2.5). This naturally implies out-of-band signaling. It also implies that call control can be exerted over several calls, or over the switch itself. The call control functions a third-party CTI application could exert are similar to those a human attendant could exert using a specialized telephone set with enhanced privileges, such as an operator's console.

In the third-party CTI application, the computer, the keyboard and screen, and the phone have no relationship to one another unless the computer establishes one. These environments tend to use the shared media processing hardware model, and tend to perform signaling via SS7 or (more commonly)

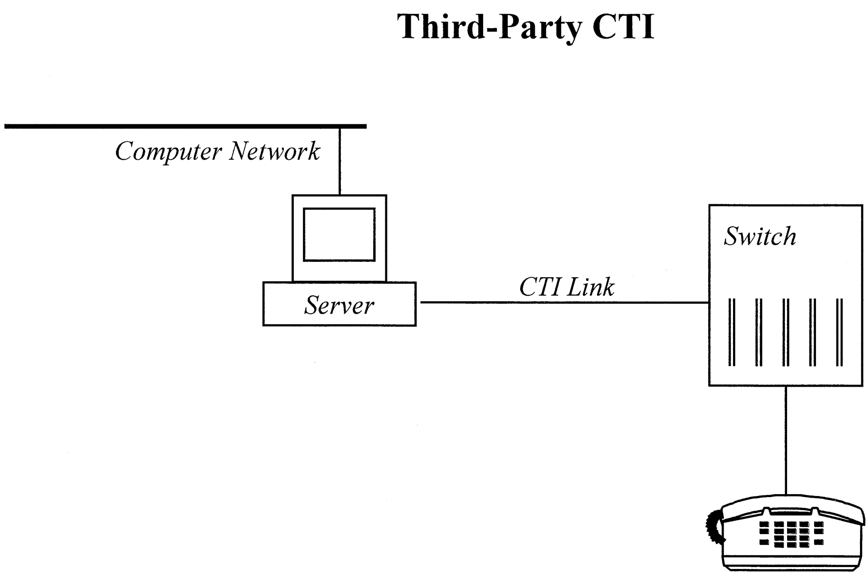

FIGURE 1.2.5 Third-party CTI model.

dedicated CTI links implementing the CSTA protocol. The CTI link typically terminates in a server rather than a specific application end-system.

There are three basic flavors of third-party CTI, which reflect the essential relationship between the computer and the switch. In the *compeer* model, the computer and switch are on equal terms. Each operates as the master of its own realm, passing information and receiving instructions from the other across a specialized interface. In the *dependent* model, the computer rules and the switch obeys. The switch has no innate call handling capability, and is actually incapable of processing calls without receiving instructions from the computer. Finally, the *primary* model is virtually identical to the compeer model, but the computer and switch do not share a specialized link. Rather, the computer attaches via a standard trunk or line port. Over the years, the dependent and primary models have seen diminishing emphasis as the market moves toward the compeer model. Unless explicitly identified as dependent or primary, third-party CTI is usually assumed to operate on the compeer model.

Automatic call routing applications are classic examples of third-party CTI. A server-based application is alerted, by the switch, to the arrival of a call. Based on ANI information, or the specific DNIS (i.e., called number), the computer directs the switch to divert the call to a specific line.

As with first-party CTI, the first third-party APIs were developed by manufacturers to support applications running on their own systems. Examples included the CallPath API from IBM, and the Computer-Integrated Telephony (CIT) API from Digital Equipment Corporation (DEC). Unlike the Hayes command set, however, none of these have achieved de facto standard status.

In the 1990s, three major APIs emerged, all strongly associated with a particular computing environment. Novell (www.novell.com) and Lucent collaborated to create the Telephony Services API (TSAPI). Novell's commercial product based on TSAPI is called NetWare Telephony Services, which links applications on remote clients with telephone system driver modules. TSAPI defines the boundary between CTI application software, and the drivers that control the links and signaling into the network.

Microsoft (www.microsoft.com) and Intel collaborated to create the Telephony API (TAPI). Like TSAPI, TAPI is concerned with call control. However, the TAPI architecture actually defines two distinct interfaces (see Figure 1.2.6). The first interface resides between CTI applications and the Windows operating system (OS). This interface, which unfortunately has the same name as the overall architecture, provides a standard means for CTI applications to access the telephony services provided by the Windows OS.

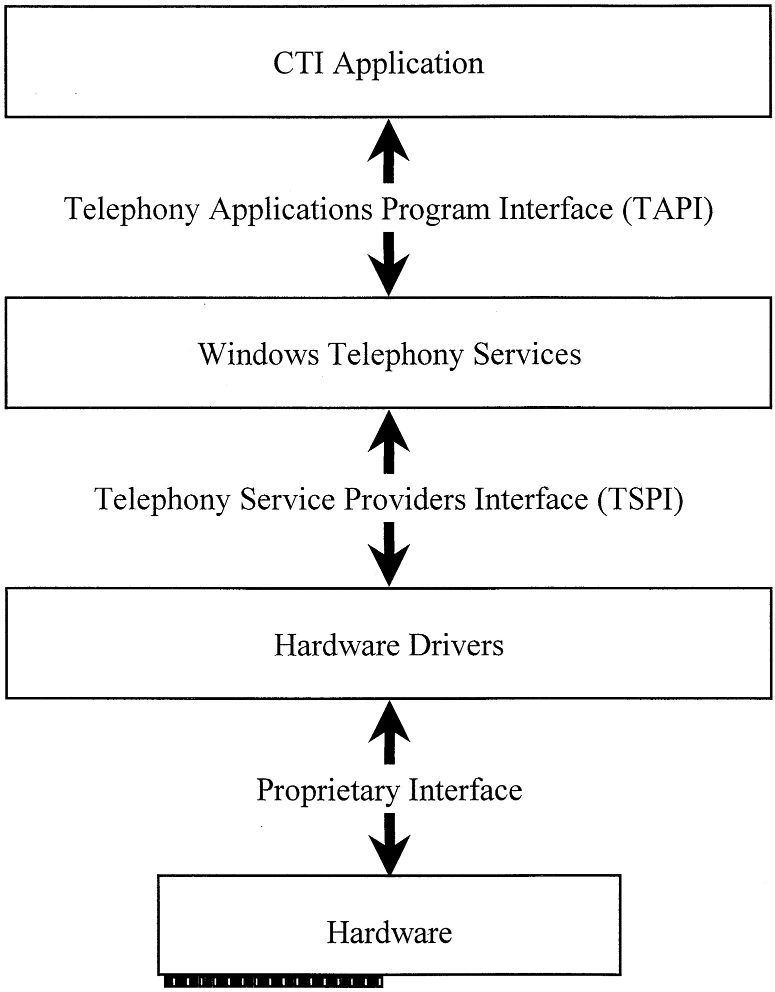

FIGURE 1.2.6 The TAPI architecture.

The second interface resides between the Windows OS and the CTI hardware drivers. Known as the telephony service providers interface (TSPI), this interface provides a standard mechanism for hardware vendors to write drivers that can support the telephony services provided by Windows. It is Microsoft's job to ensure that TAPI-compliant applications can access all of the resources provided by TSPI-compliant hardware drivers.

The third call control API is the more recent, introduced in October 1996, and brings CTI into the world of the Internet and the World Wide Web (WWW). Developed jointly by design teams from Sun, IBM, Intel, Lucent, Nortel, and Novell, the Java Telephony API (JTAPI) defines a call control interface for CTI applications running as Java applets. This opens the door to creating Web-based CTI applications. The Sun Microsystems product that implements this API is called JavaTel™.

Figure 1.2.7 integrates the various standards and concepts introduced in this paper into a single CTI model. A CTI application can be either first-party or third-party. First-party applications tend to use local, proprietary APIs (e.g., the Windows APIs) to access local call control and media processing services, and the Hayes command set to control dedicated telephony hardware.

Third-party CTI applications tend to use sophisticated call control APIs like TAPI, TSAPI, or JTAPI, and standardized media processing APIs like those defined by the ECTF. The link between the CTI server

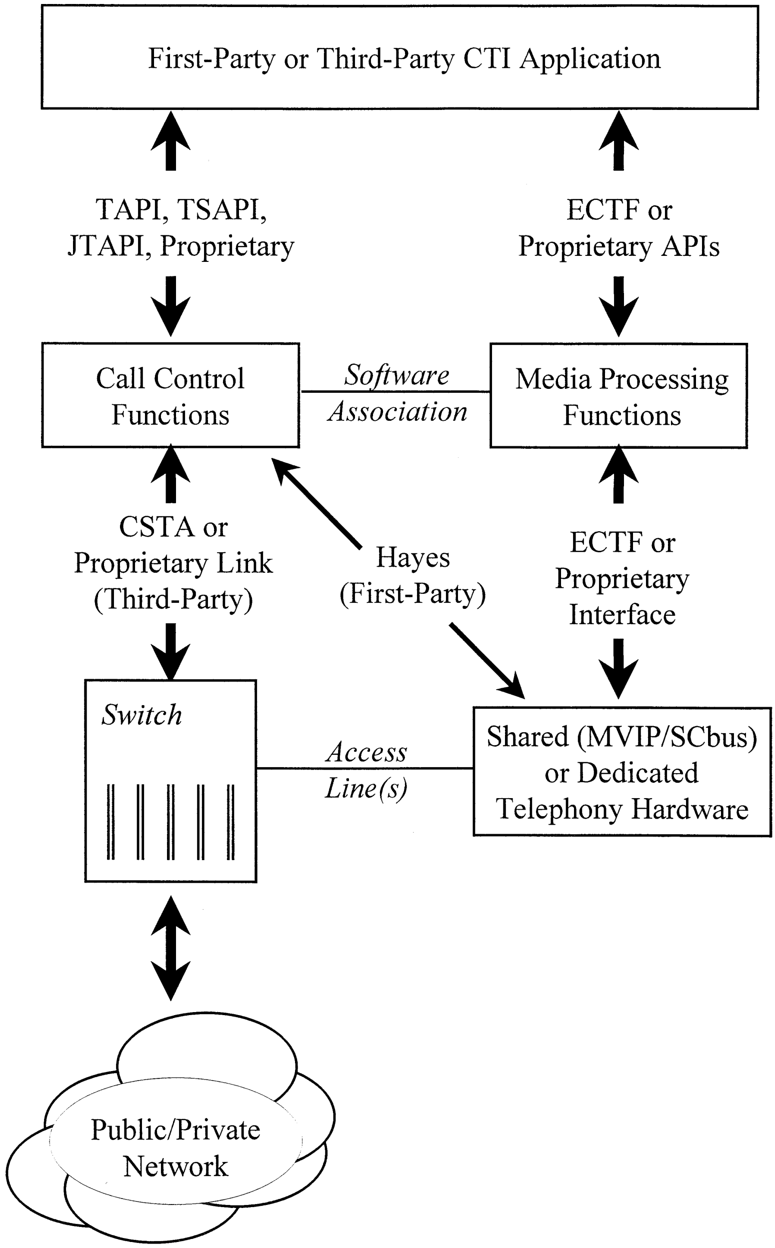

FIGURE 1.2.7 Combining the standards and components.

and the switch commonly implements the CSTA protocols. The server typically uses shared telephony hardware that is interconnected using the MVIP or SCbus architecture.

It is also possible to build a CTI server that supports several APIs and standards simultaneously. Such a product would have to map requests from all APIs into a single common function set. Dialogic's CT-Connect product takes this approach. It supports both the TAPI and TSAPI interfaces and includes built-in drivers for the ECMA CSTA link protocol and several other proprietary CTI link protocols.

1.2.5 CTI Applications and Trends

A few of the more common, and simpler, CTI applications have already been noted: voice mail, autocall, and automatic attendant. Each of these is commonly implemented as first-party CTI applications using dedicated media processing hardware. Digital dictation is another CTI application that is virtually identical to voice mail, but typically supports longer record times. The recorded dictation is usually retrieved and transcribed locally.

Many companies are beginning to provide interactive or on-demand fax services. For example, the real estate company could provide automated faxes of current properties for sale. In such a service, the user dials in and, using a touch-tone driven menu system, requests a particular fax or group of faxes and provides the number to which the fax is to be sent. The service retrieves the fax from a local file, initiates an outbound call to the specified number, and transmits the fax. As with the automated attendant application, interactive fax could be implemented as a first-party of third-party application.

Many pay-per-call applications are CTI applications. This is a common strategy for implementing fee-for-access Internet services. The user dials a 900 number and the PBX routes the call to the CTI application. The user is prompted to provide a code identifying the service they are trying to access. The CTI application provides an access code that permits the user to access the web site. The phone service bills the user for the 900 call and passes the majority of the fee to the pay-per-call service provider. The pay-per-call service provider takes an additional cut and passes the remainder of the fee to the company hosting the web service.

Perhaps the most common third-party CTI application is the inbound and outbound call center. Inbound call centers typically integrate an automatic attendant to collect initial customer information (i.e., credit card numbers, zip codes, pin numbers, etc.) and provide core services (e.g., account balances, mailing addresses, account histories, a list of service or product options, automated order taking, etc.). The caller always has the option, however, to abandon the automated system and speak to a person. In this case, the CTI application routes the call to an available attendant and provides all information the user has submitted. The application may also provide any call information provided by the phone network and any customer data retrieved from the computer's database.

The CTI market is showing clear signs of accelerated growth, fueled by a number of enabling factors in the industry. The pervasive deployment of LANs and internetworks provides the infrastructure over which many first-party and third-party CTI applications operate. The growth in digital communications and integrated networks that provide enhanced signaling capabilities (e.g., ISDN and digital telephones) create a rich set of network information on which CTI applications can be built.

The emergence of standard APIs in both the media processing and call control arenas has furthered equipment and service interoperability. Furthermore, the increasing maturity of voice processing technology makes interactive voice response (IVR) systems easier to deploy and use. Finally, the industry is seeing a broad array of CTI application development toolkits. Examples of these include OmniVox from Apex Voice Communications (www.apexvoice.com), Visual Voice from Artisoft (www.artisoft.com), MasterVox from Mastermind Technologies (www.mastermind-tech.com), and IVS Builder and IVS Server from Mediasoft Telecom (www.mediasoft.com).

1.2.6 Conclusion

The CTI market is a young one, but the technologies coming together into this application environment are relatively mature. As the CTI-related standards themselves mature, interoperability agreements emerge, and economies of scale begin to apply, CTI applications are likely to become pervasive. Furthermore, with the emergence of JTAPI and the increasing drive toward voice over IP (and hence over the Internet), CTI applications are finding a new niche in which to grow. The Internet is a significant niche indeed!

For further information, the reader is recommended to visit the various web sites identified in this chapter. There are also two periodical publications dedicated to CTI, both of which can be accessed via the Internet: *Computer Telephony* (www.computertelephony.com) and *CTI Magazine* (www.tmcnet.com).

1.3 Voice over IP

Matthew Kolon

1.3.1 The Coming Integration of Voice and IP Data

Companies in the U.S. spend $100B on long-distance and international telephony every year. Most of that money goes to the basic transit of voice and fax from one location to another. With the continued pervasiveness of intelligent peripheral (IP) networking, a new class of products and services has evolved to move some of that traffic from its traditional home on the public switched telephone network (PSTN) to a variety of packet-switched networks. While many of these new "voice" networks have not previously been considered telephony-class, they are nonetheless attractive because of their low cost.

The IP telephony scene has jumped from being a hobbyist's realm of custom solutions and cobbled-together software to a $400M per year industry hotly pursued by industry giants of hardware and software. Continued improvements in digital signal processor (DSP) technology, voice packetization techniques, and the networks that IP voice runs over have combined to make the start of the 21st century into the era that IP telephony begins the transition to a mainstream solution for business.

There are a number of reasons for the inevitability of this transformation, but all of them come back to the relief of high-cost long-distance telephone services. Reviewing a few comparative facts regarding the PSTN and voice over IP (VoIP) presents some compelling realities:

- **One can fit more voice on an IP network than one can on the PSTN.** The Bell System definition of a single voice channel as a 64kbps DS-0 has led to a long-standing institutional belief that 64k is necessary to carry a voice conversation. Thus a T-1 is commonly referred to supporting 23 "voice" channels over its 1.544 Mbps. Yet today's VoIP products can carry hundreds of voice conversations over that same amount of unchannelized bandwidth.
- **Packet networks are much better than they used to be.** Improvements in the quality of physical-layer packet networks over the past 30 years have resulted in a large general improvement in data integrity. The same forces that make simple frame relay an effective replacement for the robust X.25 protocol mean that even connectionless IP data — and voice — may be entrusted to today's connectionless networks and still have an excellent chance of getting through in a reasonable amount of time and with few errors (or little delay) of consequence.
- **Control of IP data networks rests largely in the hands of the customer.** As long as a minimum quality of service — particularly the establishment of maximum delay guidelines — is met, virtually every service available over IP is controllable from the sending and receiving stations. For example, packets may be routed over the Internet for free if tolerant of lower quality, over a private IP network if demanding of higher quality, or even over the PSTN if necessary — all at the discretion of the originating node.

These are just a few of the reasons why many network managers are examining the current possibilities for placing at least some of their voice traffic into IP networks.

1.3.2 Applications for Voice over IP (VoIP)

Of course, with long-distance services being the single most expensive portion of any company's telephony budget, the application of VoIP to the interexchange carrier (IEC) realm is taking the forefront when it comes to the immediate application of the technology. The basic design of such a network is rather simple: gateways within local calling areas connected by an IP network which spans the distance previously covered by the IEC.

While a company implementing VoIP for the purpose of saving charges on interoffice communications may have a design as simple as that in Figure 1.3.1, it is more likely that the IP network will connect multiple sites, each with its own gateway, each of which may then contact another dynamically when it

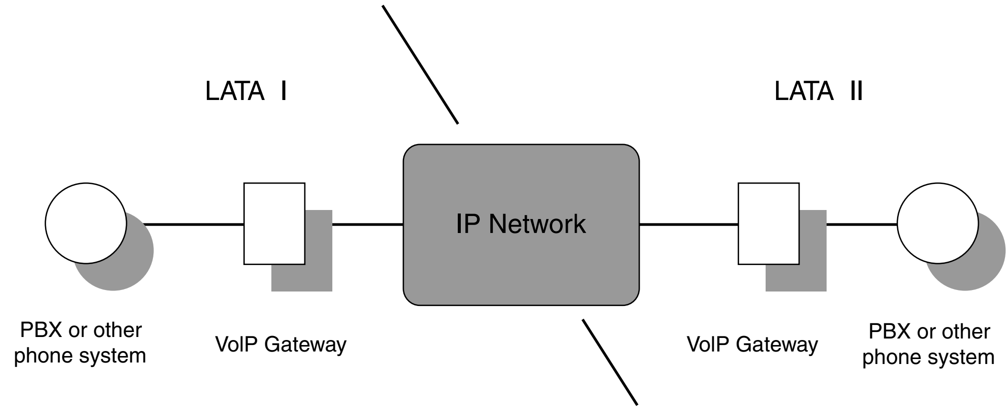

FIGURE 1.3.1 Business IEC replacement using VoIP.

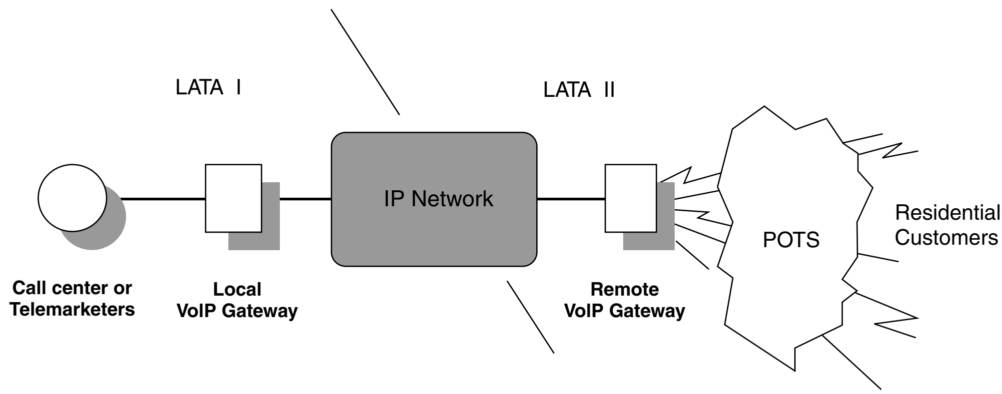

FIGURE 1.3.2 Business-to-residential VoIP network.

has a voice call destined for that site. The connectionless nature of IP ensures that new gateways may be added at will, with little need for reconfiguration at the other stations.

Many variations of this scheme are possible, depending upon the nature of the service one is trying to implement. For tie-line replacement and business-to-business calls, the simplest to exploit is that shown in Figure 1.3.1, that is, two or more gateways connected by an IP network. The reason that most pundits consider this setup to be the first area to exploit VoIP is because the difficult part — getting the voice to a few places where it can be digitized and packetized into IP — is already done. The private branch exchange (PBX) that currently connects via a leased line or IEC to another PBX can easily have that connection replaced by IP — with no changes in how users place calls.

Another application that is generating a large amount of industry interest is that of business-to residential telephony (Figure 1.3.2), to allow telemarketers or call centers to physically centralize while obtaining low-cost long-distance service via VoIP. In this scenario, residential customers are able to dial a local number and access a VoIP gateway which connects them to the implementer's customer support or sales office — wherever it may be. The customer makes a free call, and receives the same service had an 800 number been dialed, but the company avoids the cost of maintaining 800 service. It is also able to supply customers with a "local" number to call for service, which can enhance the company's image.

Reversing the above strategy — that is, using the remote gateway to *place* local calls rather than accept them — allows telemarketers access to large, yet distant, markets without the need to place large numbers of long-distance calls to get to them.

Yet another option exists for those eager to exploit the possibility of VoIP at their businesses or campus: replacing the PBX and its network with an IP network. Most businesses are already halfway there; they have local area networks (LANs), routers, and digital wide area network (WAN) facilities capable of handling IP traffic. New products, such as 100- and 1000-Mbps Ethernet, as well as the cost-effective speed of LAN switching, mean that network managers can build an enormous amount of capacity into their local and enterprise networks — capacity which might well be used to carry voice traffic. Traditional models for business traffic have always involved the creation and management of two separate networks, one for voice and one for data. The encapsulation of voice in IP packets means that the consolidation of voice into the data network is now possible, with the corresponding reduction in the need for equipment, data facilities, staffing, and expertise in several types of systems. Consolidation of voice traffic and data traffic into the same end-to-end network opens the door to true integration of messaging and telephony systems, such as integrated email and voice mail, and IP-based fax messaging.

The final area of interest for VoIP proponents is that of residential-to-residential connectivity, that is, friends and relatives speaking to each other from handsets or speakerphones integrated into Internet-connected PCs. While this is the application that "proved" the possibility of VoIP, it remains the most difficult to ensure acceptable quality for. The difficulty of obtaining quality voice this way has nothing to do with the equipment at the ends of the link, but rather with the lack of guaranteed, or even reliable, values for delay and delay variation over the Internet. Indeed, improvements in low-cost digitization hardware and "Internet telephony" software have made it possible to have a full-featured, high-quality VoIP gateway for the cost of a new PC. But even the best-quality digital voice will be unintelligible if only half of it arrives at the intended destination.

These are just the basic categories that some of the most obvious applications for VoIP fall into. But applications are as numerous as those for the telephone itself — perhaps even more so. The lower cost of VoIP means that some uses for telephony that were once deemed uneconomical may now be justified. And the integration of voice and data traffic over a single IP network may make some forms of integration possible that were unthinkable just a few years ago.

1.3.3 A Component-based Overview

What are the components of a successful IP telephony system? While there are of course a number of different approaches, there are a few basic ingredients that all systems must implement — although the use and location of parts changes with different network designs.

The VoIP Network: In the list of VoIP components (Figure 1.3.3), the IP network(s) over which the voice will travel is of primary importance. IP is first and fundamentally a connectionless protocol, with no guarantees concerning the traffic that it carries. It cannot ensure a maximum delay or variability of delay, cannot retransmit errored or lost packets, and does not even promise that its payload will arrive at all. The quality of service one receives from the PSTN, and that provided by even the most carefully managed and overbuilt IP network do not bear comparison. And for those thinking about using the Internet as the equivalent of their current expensive IEC service…well, suffice it to say that when a web page often takes 60 seconds to download, sending real-time voice traffic over that same series of links will be a challenge. Until the Internet infrastructure is managed under an agreement which includes concrete plans to provide some limited and predictable delay — in an interprovider fashion — voice traffic cannot travel the Internet and maintain the quality that business customers demand. It's worth mentioning that this agreement is nowhere in sight.

That does not mean that today's Internet has no place in the voice network, however. VoIP gateways can use the Internet to provide the non-real-time services that constitute much of today's "voice" traffic. The most obvious one of these is facsimile transmission. While fax machines thrive on the dedicated lines of the circuit-switched PSTN, there is no reason why their transmissions cannot be placed in IP for long-distance transit. Delay — the reason why interactive voice is so difficult over the Internet — doesn't affect fax transmissions at all, and transmission control protocol/Internet protocol (TCP/IP) can resend

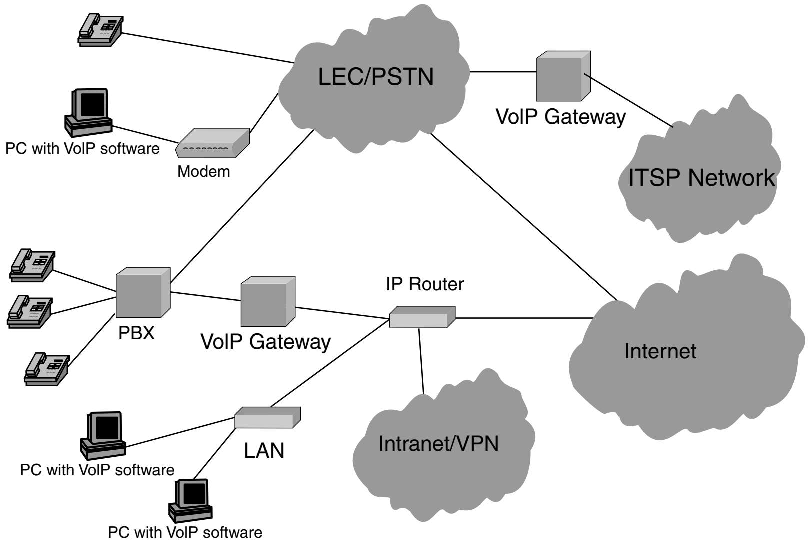

FIGURE 1.3.3 VoIP network components.

data until the network gets it right without bothering the receiver. The same could be said for voice mail messages.

The next step between the very public Internet and a completely private IP network is the ISP backbone itself, which is nothing more than a single provider's portion of the Internet. If this network extends close to the points where gateways will be placed, IP traffic between them may remain solely on that network. In almost all circumstances, this will result in less delay and better predictability for traffic of all types. But while the statistics for network performance may improve in a single-provider environment, the lack of user control over these fundamentally public networks may be unacceptable for the network manager who seeks to have some influence over the environment in which his traffic travels. Single Internet service provider (ISP) IP telephony, though, has the lowest cost of any of the non-Internet options, and therefore is attractive as long as acceptable quality can be achieved. This may be a matter of simply trialing a number of ISP networks and choosing the one with the best performance, or may actually involve a level of performance — with stated delay and throughput characteristics — to be specified in the user contract.

Luckily, the Internet and its constituent networks are not the only options for long-distance carriage of VoIP. Many of the larger ISPs offer, in addition to their public Internet network, access to a separate IP network designed for virtual private network (VPN), intranet, extranet, and other semiprivate usage. These networks are not any more remarkable in concept than an average ISP's network, except for their managed nature, that is, the knowledge the provider has of just how much traffic any one user is likely (or allowed) to subject the network to at any one time — something unheard of on an Internet access network. This knowledge allows the provider to predict and maintain a high level of quality, which can result in service level agreements in which end-to-end delay is specified to be well below 0.5 seconds — the point at which telephony starts becoming reasonable. In this environment, SLAs are becoming the rule rather than the exception.

The ultimate VoIP network, however, is the one where all aspects of IP traffic and performance can be managed by the users — a completely private intranet. Formed from private (leased) lines, with perhaps some links composed of frame relay or asynchronous traffic mode (ATM), the distinguishing characteristic of these networks is that they are completely under the control of the network managers who deploy and run them. Therefore, the amount of bandwidth reserved for voice traffic can be strictly controlled, as can the throughput of routers and other connectivity equipment. How those resources are actually apportioned may vary from protocol-based reservation systems like reservation protocol (RSVP) to completely manual intervention, but whatever the method, the manager has the ability to restrict the effect of data traffic that interferes with voice. While this sounds like — and in fact is — the ideal environment for packetized voice, it comes with a price. Completely private IP networks are by far the most expensive way to ship IP from one location to another. Whether the establishment of such a network is worth the ability to carry voice effectively depends on how much money can be saved by eliminating IEC charges from the IT budget.

If the number of options and the headaches of managing another network service are a serious disincentive, another possibility is to leave the network and its management to the specialists — that is, to contract with one of the growing number of Internet (or IP) telephony service providers (ITSPs). An ITSP functions as a plug-and-play replacement for a traditional IEC, by providing the gateway, network, and management needed to make VoIP successful. The tradeoff here, of course, is that since the ITSP does all the work, they also reap some of the rewards. Typically, ITSPs function like an IEC in terms of billing, with per-minute rates that range from one half to three quarters that of comparative IECs.

That level of discount may change before long, however. Much of the savings that ITSPs are able to pass on to their customers are possible because of a May 1997 FCC ruling that classifies ISPs and ITSPs as end users of the PSTN rather than as carriers. This classification currently makes it impossible for LECs to charge ITSPs the same access charges they demand from traditional IECs. Those access charges, when passed on to the IEC customer, can account for as much as one half of the average IEC bill. It is the lack of these charges, more than the technological benefits of VoIP, that allows ITSPs to sell services for so much less than their IEC counterparts.

While the level of savings on recurring charges is the least with the ITSP option, it may well be compensated for by the simplicity of setup and management, and the lack of gateway hardware or software costs. The users who benefit from the access charge loophole, however, may have some hard decisions to make if, as many believe will occur, the FCC reverses itself and decides to consider ITSPs as carriers. In that market, much of the price differential would disappear, and users would have to make their decisions based more on quality, service, and other points rather than price (Figure 1.3.4).

All of these networks can and will benefit from work currently underway to allow efficient prioritization of packets containing voice over those containing non-real-time data. Gigabit-speed routers, faster switches, better routing and path-reservation protocols, and the continued addition of cheap bandwidth are all reasons why VoIP quality will continue to increase.

In summary, there are a number of network options for VoIP. Which one best suits a particular need depends on a number of factors, primarily revolving around the level of expected quality. For those looking for a way to lower the cost of interoffice communications — an application where the "internal" aspect may allow slightly lower quality than that required for communications with customers — some of the lower-cost options like single-ISP VoIP networking may suffice. Those wishing to completely replace their IEC contract with an IP-based IEC solution are faced with replacing a complex network from the ground up, and will have to plan, and pay for, a much more robust service. And for the time

Network	Gateway	Cost	User Control	Performance
Internet	User-provided	Least	Least	Worst
Single ISP	User-provided			
Managed IP	User-provided			
Private IP	User-provided		Most	Best
ITSP	Included in contract	Most	N/A	N/A

FIGURE 1.3.4 VoIP network options compared.

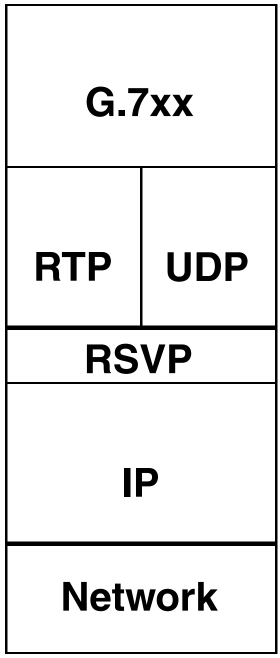

FIGURE 1.3.5 VoIP protocol components.

being, at least, voice over the public Internet remains in the realm of a hobby for those willing to tolerate indifferent and completely unpredictable voice quality.

Gateway Software and Hardware: The hard work of actually taking analog voice and sending it over an IP network, as well as receiving IP and converting it back into voice, is the job of the gateway. It is easiest to examine the issues related to this complex task if we break it down into its components (Figure 1.3.5):

Accept analog or digital voice: A gateway must have some connection to the non-IP world where the voice traffic originates, usually consisting of either a bank of dial-in plain old telephone service (POTS) ports or a digital connection to a PBX.

Prepare the voice signal: In order to use the available bandwidth as efficiently as possible, the voice signal must go through a number of transformations before it is ready to be digitized. First, it must be "cleaned up" by having as much noise and echo removed as possible. The techniques for doing this have been well established in the traditional telephony world for years, but the cooperation of the various systems and gateways through which voice may pass is essential. This means that calls traveling through a LEC on their way to the VoIP gateway may need to be treated differently than those coming directly from a PBX.

Second, it must be stripped of unnecessary silence, to avoid making the gateway send hundreds or thousands of packets per second carrying nothing. Most gateways have adjustable options for when silence suppression "closes off" and stops transmitting on behalf of a user, but the effectiveness of default settings may depend on usage characteristics that are themselves dependent on cultural factors. Some adjustment of this setting to achieve the best compromise between quality and throughput is usually necessary. Related to the subject of silence suppression is the modeling and regeneration (at the remote end) of background noise, without which users can become disconcerted.

Compress and digitize the voice signal: The standard compression and digitization of voice provided by traditional 64k PCM produces a stream of digital data that is enormous compared to that available by many newer codecs. While some vendors have achieved good results with proprietary schemes, most of the industry is settling down to the use of one or another International Telecommunications Union (ITU) G-series codecs, as specified in their H.323 standard. H.323 is a complex specification for point-to-point and multipoint teleconferencing, data sharing, and telephony over IP. While the full effect of this standard on "VoIP-only" products remains to be seen, the G.711, G.723, and G.729 codec specifications referenced by it are current favorites for coding voice.

These three standards differ primarily in the amount of work that the DSP must do in order to process the analog signal, and the number of bits that it takes to represent a given amount of voice. While recent advances in DSP design and manufacture have allowed vast improvement in these areas, there remains an inverse relationship between them, and also therefore a higher cost for greater efficiency. Nevertheless, the most aggressive of the standards — G.729 — can represent 10 msecs of voice with only 10 octets of IP data. The less intensive G.711 and G.723 trade higher traffic volume for higher quality. Many gateways can be configured to use whichever one of these standards provides the most acceptable trade-off between quality and traffic level.

Route the call: Once a gateway has a potential stream of packets ready to send, it must have some way to identify the address of the gateway it will send them to, and to inform that gateway of which local user it is destined for (or what local number to dial.) For simple point-to-point applications, IP address can be a manually configured variable, since there is only one destination possible. But in cases where a multipoint network means that packets may be simultaneously distributed among a number of destinations, there must be a process in which the called number is translated into an IP address.

Informing the destination gateway of the called phone number has its complications, too, because many of the codecs used in current gateways compress the analog signal so much that the dual-tone multi-frequency (DTMF) tones produced by phones become unreliable. Therefore, the calling gateway must be able to transform those DTMF tones into a code representing the called number and transmit them to the destination gateway for correct routing at the called end.

Packetize and send digital voice in IP datagrams: At first glance, this is the simple part. After all, IP stacks on end stations and routers have been performing this function since the late 1960s. Yet some of the characteristics of packet-switched networks with regard to real-time traffic are different than those regarded as common knowledge by those used to thinking of IP as data-only transport.

For example, the flexible size of an IP datagram, while an advantage in the transmission of data, complicates the problem of achieving low variability of delay, since IP routers handle packets of various sizes differently, and may tend to process smaller packets more quickly than larger ones. The destination gateway would then need to account for the tendency of larger packets to take longer, and thus delay reassembly. In practice, VoIP gateways by default transmit packets of a single size or small range of sizes in order to obviate this problem, but this is one area where the capabilities of the gateways and the network(s) over which they will transmit must be closely matched. Setting the maximum packet size of the gateway to any amount higher than the maximum transmission unit (MTU) of the underlying network will introduce latency as routers fragment datagrams that are too big to travel through networks attached to them.

Enabling routers to prioritize packets containing voice can enable voice and data to coexist on the same network more easily. Methods for doing this include enabling priority queuing based on transport layer port number, packet size, and source and destination addresses. RSVP can be used to reserve router bandwidth and processing capability, as well as network segment bandwidth, for packets that meet certain criteria, but implementing RSVP demands a network path in which all routers are RSVP-compliant, something that is not likely in a multiprovider (or even some single-provider) scenarios.

Receive, buffer, and decode the incoming stream of VoIP data: Again this is a well-understood process for data, which generally depends upon the IP suite's TCP protocol to retransmit lost data and reassemble segments in the proper sequence before it is passed to the application. VoIP software seldom makes use of TCP, largely because the services it provides introduce far too much latency into the transmission process for them to be useful (an exception to this rule is fax transmission, for which TCP makes sense given the lack of need for real-time treatment of data.) Instead, most gateways can use real time protocol (RTP) as the protocol in which voice data rides. While having no control over delay imposed by the network, RTP makes it possible to trade a small amount of additional delay for a reduction in the amount of delay variation. This is accomplished by transmitting each packet with a timestamp that can be read by the receiver and used to pass data to the upper layers of the VoIP software with something like the transmitted amount of inter-packet delay.

Alternatively, some gateways have the option to send digitized voice in user datagram protocol (UDP) packets, which travel in an unstructured stream, free of sequence numbers, timestamps, and acknowledgments — but also free of the delay imposed by processing these variables. Since the audio stream at the remote end must go on regardless of the actual receipt of data, large numbers of packets that are lost en route simply result in "holes" or "dropouts" in the audio signal. While this sounds as though it would spell the end for reproduction of any reasonable quality, in fact it takes the loss of a relatively large number of packets to create noticeable holes in outbound audio at anything but the highest compression levels. Whether the control and complexity of RTP or the simplicity and speed of UDP will prove to be the most effective way to carry datagram voice remains to be seen.

1.3.4 Keys to Successful Deployment

The large number of configurable variables and the many options within each make configuring VoIP networks a considerable challenge, especially since these networks' main role is to replace some of the most bulletproof networks in the world: those of the PSTN. Aside from performance issues, questions of interoperability abound, particularly for those users who wish to deploy distributed VoIP networks consisting of hardware and software from more than one vendor, and networks from more than one provider.

One thing is certain, though: IP telephony is here to stay. Despite the challenges that network managers face in order to reduce their IEC bills, in at least some applications the payoff is great enough to make the decision to at least trial the technology obvious. The astute manager, however, remembers a few things:

- Few, if any, of the products currently available for VoIP networking work well "out of the box." Nearly everyone who has implemented gateways on either a point-to-point or multipoint basis has a story to tell about the setup and configuration of their system, and the shakedown and subsequent adjustments, that had to occur before the network settled down. Almost as invariably, though, they can recount the time that things began to work well, and now can point to users who are happy with the price and performance of the VoIP network.
- All VoIP products aren't the same. Vendors are scrambling to improve quality and add features, and that translates into large variations in product lines — at least until the next revision is introduced.

The good news is that there are many positive signs for those considering putting their trust into VoIP. The current standards situation for components of VoIP products seems to be stabilizing. While any emerging technology — especially ones with such high visibility — generates a large number of proprietary solutions which get narrowed down by the market, VoIP is one example of how vendors can cooperate. Most of the standards for encoding (the ITU G-series) seem to be settling down for a long period of maturity.

With regard to the network technologies in use, a new generation of network designers and engineers feels more comfortable with IP than with any other technology — including voice traffic. The ubiquity of the Internet and of IP itself have created a large pool of experience from which managers can draw when deploying VoIP. As for the future, a knowledge of the workings of Internet protocols is commonplace among graduates of almost any technical program.

While the public telephone network has existed for years, fast public data networks have not existed until recently, and new data networks are being constructed at a staggering rate. Many of these networks will be suitable for voice traffic, and thus can extend the reach of VoIP networking. And the rapid pace of network improvement means that end-to-end latency will continue to drop, which can only mean good things for the quality, and success, of VoIP.

Acronyms

ATM — Asynchronous transfer mode
DSP — Digital signal processor
DTMF — Dual-tone multi-frequency
FCC — Federal communications commission
IEC — Interexchange carrier
IETF — Internet engineering task force
IP — Internet protocol
IP — Intelligent peripheral
ITSP — Internet (IP) telephony service provider
LAN — Local area network
LEC — Local exchange carrier
PBX — Private branch exchange
PSTN — Public switched telephone network
RSVP — Reservation protocol
RTP — Realtime protocol
UDP — User datagram protocol
VoIP — Voice over IP
WAN — Wide area network

1.4 Local Area Networks

John Amoss

1.4.1 Overview

1.4.1.1 Standards

The Institute of Electrical and Electronics Engineers (IEEE) 802 Local and Metropolitan Area Network Standards Committee has the basic charter to create, maintain, and encourage the use of standards for local and metropolitan networks. In the IEEE 802 Committee context the term "local" implies a campus-wide network and the term "metropolitan" implies intracity networks. The IEEE 802 Committee defines interface and protocol specifications for access methods for various Local Area Network (LAN) and Metropolitan Area Network (MAN) technologies and topologies. The project has had a significant impact on the size and structure of the LAN market.

The standards are jointly published by the IEEE, the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC). An overview of the standards is published by these bodies. [1,2]

1.4.1.2 Reference Model

Figure 1.4.1 relates the specific protocol layers defined by the IEEE 802 Committee, which include Physical, Media Access Control (MAC) and Logical Link Control (LLC) layers, to the layers of the Open Systems Interconnection (OSI) Reference Model. [3] The protocol architecture shown in Figure 1.4.1, including the Physical, MAC and LLC layers, is generally referred to as the IEEE 802 Reference Model.

Working from the bottom up, the Physical layer of the IEEE 802 Reference Model corresponds to the Physical layer of the OSI Reference Model and includes the following functions.

- Encoding/decoding the signals to be transmitted in a manner appropriate for the particular medium, e.g., the use of Manchester or Non-return to Zero encoding schemes;
- Achievement of synchronization, e.g., by the addition of a preamble field at the beginning of a data frame;

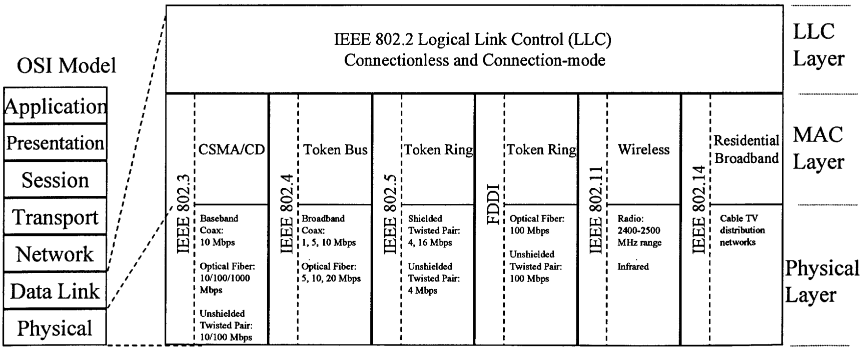

FIGURE 1.4.1 IEEE 802 reference model.

- Bit transmission and reception;
- Specification of the physical and electro/optical characteristics of the transmission media (e.g., fiber, twisted pair wire); and
- Network topology (e.g., bus, ring).

Above the Physical layer are functions concerned with providing the frame transmission service to LAN users. Such functions include the following.

- Governing access to the LAN transmission medium.
- Performing error detection (e.g., via addition of a Frame Check Sequence field);
- Assembling the frame for transmission; and
- Upon reception, performing address recognition.

These functions are collectively associated with a MAC sublayer, shown in Figure 1.4.1. As indicated in the figure, a number of MAC layers are defined within the IEEE 802 Reference Model including access control techniques such as Carrier Sense Multiple Access/Collision Detection (CSMA/CD) — also generally referred to as Ethernet — Token Bus and Token Ring.

Finally, the Logical Link Control (LLC) layer is responsible for providing services to the higher layers regardless of media type or access control method (such as those specified for CSMA/CD, Token Bus, Token Ring, and so on). The LLC layer provides a High-level Data Link Control (HDLC)-like interface to the higher layers and essentially hides the details of the many MAC schemes shown in Figure 1.4.1 from the higher layers. The LLC layer provides a multiplexing function, supporting multiple connections, each specified by an associated destination service access point (DSAP) and source service access point (SSAP), discussed later. As shown in Figure 1.4.1, the LLC layer provides both connectionless and connection-oriented services, depending on the needs of the higher layers.

1.4.1.3 Overview of the Major MAC Standards

Since its inception at Xerox Corporation in the early 1970s, the carrier sense multiple access with collision detection (CSMA/CD) method, also commonly termed Ethernet, has been the dominant LAN access control technique. The CSMA/CD method was the first to be specified by the IEEE, under the IEEE 802.3 working group, and was closely modeled after the earlier joint Digital/Intel/Xerox (DIX) Ethernet specification. [4] Ethernet has, by far, the highest number of installed ports and provides the greatest cost performance relative to other access methods such as Token Ring, Fiber Distributed Data Interface (FDDI) and the newer Asynchronous Transfer Mode (ATM) technology. Recent and in-progress extensions to Ethernet include Fast Ethernet, which, under the auspices of the IEEE 802.3u working group, increased Ethernet speed from 10 Mbps to 100 Mbps thereby providing a simple, cost-effective option for higher speed backbone and server connectivity, and Gigabit Ethernet, which under the auspices of the 802.3z working group increased the speed to 1000 Mbps.

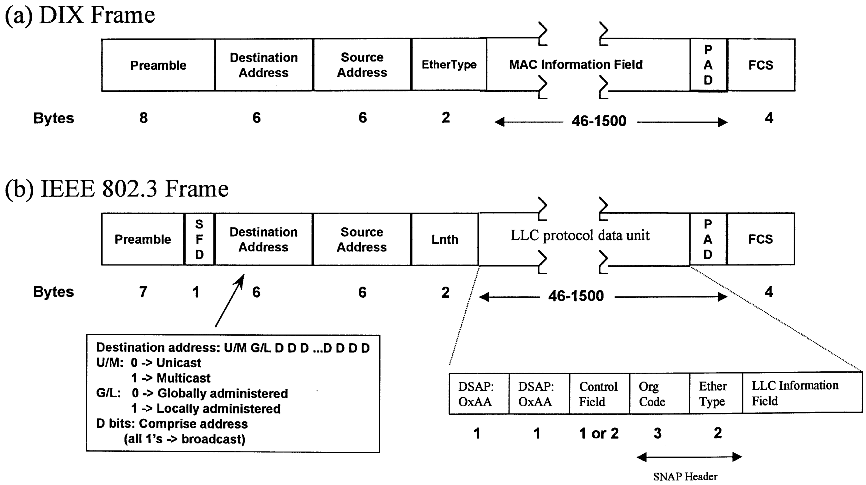

FIGURE 1.4.2 DIX and IEEE 802.3 frame formats.

The IEEE 802.4 Token Bus specifications were developed primarily in response to requirements for the deterministic performance of a token passing scheme, coupled with a bus-oriented topology. The use of a broadband technology option provided the additional benefits of increased bandwidth, geographic coverage, and number of terminations.

The IEEE 802.5 Token Ring specification was developed with major support from IBM and reflected IBM's perspective on local area networking. Improvements over the IEEE 802.3 scheme include deterministic performance and the specification of a priority mechanism.

As shown in Figure 1.4.1, work has been completed in several new technology areas including wireless LANs (IEEE 802.11) [5] and Cable Modems (IEEE 802.14). [6]

Due to their wide market acceptance, this section focuses on the details of the IEEE 802.3 (CSMA/CD) and 802.5 (Token Ring) specifications. The section also addresses the Logical Link Control layer and presents an overview of building wiring considerations which would ensure that the building cabling meets the requirements of the various LAN types.

1.4.2 IEEE 802.3 (CSMA/CD) Specifics

1.4.2.1 Frame Structure

As mentioned, the carrier sense multiple access with collision detection (CSMA/CD) method was the first to be specified by the IEEE and was closely modeled after the Digital/Intel/Xerox (DIX) Ethernet specification. Although there are differences between the Ethernet and the 802.3 specifications, manufacturers now typically produce hardware that can support both, so that effectively the two are compatible. Differences in the packet format are resolved in firmware for a particular implementation. We use the terms Ethernet and IEEE 802.3 CSMA/CD interchangeably.

The frame format in the original DIX specification is shown in Figure 1.4.2(a). Frame fields are as follows.

- Preamble — To allow synchronization by the receiving station and to indicate the start of frame, the frame starts with an eight byte sequence, the first seven of which have the format (10101010), and the eighth the format (10101011).
- Source and destination addresses are 48 bits each (a little-used option allows for 16 bits) and have the structure shown in Figure 1.4.2(b) except for a minor variation in the second bit of the address.
- EtherType — The EtherType field (16 bits) allows for the multiplexing of data streams from different higher level protocols and identifies the particular higher level protocol data steam carried

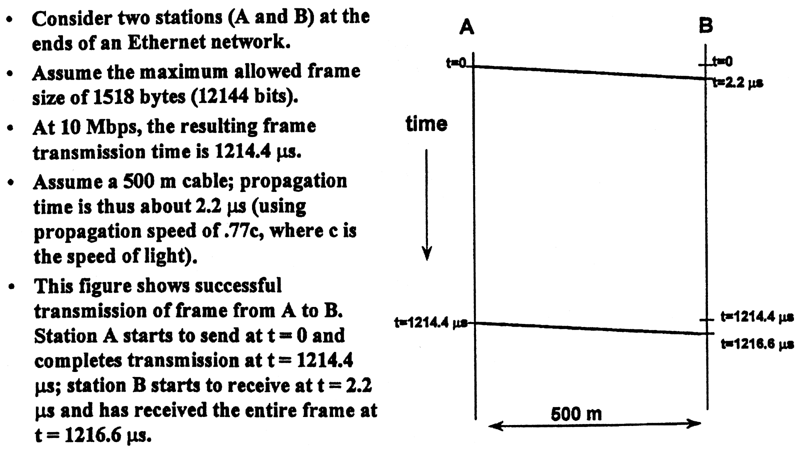

FIGURE 1.4.3 Example of successful frame transmission.

by this frame, e.g., an EtherType of Ox08-00[1] indicates a frame carrying an IP datagram. Values for the EtherType field can be found in [7].

- Data — The Data field carries the service data unit from the higher layer protocol entity and ranges in length from 46 (including an added PAD field if the service data unit is less than 46 bytes) to a maximum of 1500 bytes.
- Frame Check Sequence (FCS) — Finally, a four-byte FCS field is added for error detection purposes.

The IEEE 802.3 frame format is shown in Figure 1.4.2(b). The major difference in format arises from the need to accommodate other MAC specifications under the IEEE umbrella which may have no equivalent of the EtherType field. As a result, this multiplexing capability is included in the next higher layer of the IEEE 802 Reference Model, the LLC layer (see Figure 1.4.1). The method used to provide this additional protocol information is the Subnetwork Access Protocol (SNAP). A SNAP encapsulation is indicated by the LLC layer SSAP and DSAP fields both being set to OxAA. The SNAP header is five bytes long: the first three bytes consist of an organization code, which is assigned by the IEEE; the second two bytes use the EtherType value set from the Ethernet specifications. Using this scheme, the multiplexing service afforded by the EtherType field is available at the LLC layer, independent of the individual MAC layer capabilities. Note that several layers of multiplexing are available at the LLC layer; one provided by the LLC Destination Address/Source Address fields in Figure 1.4.2(b), and the other by the LLC/SNAP fields shown in the figure (which include the EtherType field). Again, when the length of MAC layer data field is less than 46 bytes, a PAD field is added to ensure a minimum data plus PAD field length of 46 bytes. The PAD field consists of an arbitrary array of bits.

1.4.2.2 Sample Frame Transmission

For a transmission media operating at a data rate of 10 Mbps, typical of many 802.3 specifications, Figure 1.4.3 shows the successful transmission of a frame between two stations at the ends of the cable, from station A (shown on the left) to station B (shown on the right). Cable length is assumed to be 500 meters, the approximate maximum length for a number of IEEE 802.3 configurations (per Section 13 of [8]). A frame size of 1518 bytes is assumed, also the maximum as per the IEEE 802.3 specification. From the figure, station A begins transmitting at time t = 0 and some time later the leading edge of the signal

[1]This notation indicates a string of bytes (groups of eight bits) with the values of the bytes given in hexadecimal form; thus Ox08-00 represents the two bytes 00001000–00000000.

TABLE 1.4.1 Minimum Propagation Speeds for Sample Media

Media Type	Minimum Propagation Speed
Coax (10BASE5)	0.77 c
Coax (10BASE2)	0.65 c
Twisted Pair (10BASE-T)	0.585 c

appears at station B. This time is determined by the propagation speed of the signal on the particular media, with the speeds for a number of media shown in Table 1.4.1. Assuming a propagation speed of .77c, where c is the speed of light (3×10^8 m/s), yields a propagation delay of about 2.2 μs for the example in Figure 1.4.3.

The total signal transmission time, neglecting a short initial synchronization period when the preamble and start of frame delimiter are transmitted is

$$\left(1518 \text{ bytes}\right) \times \left(8 \text{ bits}/\text{bytes}\right) / 10 \text{ Mbps} = 1214.4\ \mu\text{s}$$

Thus station A completes transmitting the signal at t = 1214.4 μs and station B begins receiving the signal at t = 2.2 μs and receives the entire signal at time t = 1216.6 μs.

After a brief delay period to allow recovery time for other stations and the physical medium, termed the interframe gap, another frame can be transmitted if available. An interframe gap of 9.6 μs or 96 bit times for a 10 Mbps implementation is specified by the standard. This value is chosen to account for variability in the gap as frames travel over the media and connecting repeaters (discussed below). This variability occurs because two successive frames may experience different bit loss in their preambles. If the first packet experiences greater bit loss than the second, the gap will shrink as the repeater reconstructs the preamble and therefore introduces delay. If the second frame experiences greater bit loss, the gap will expand.

1.4.2.3 Carrier Sense Multiple Access

A simple addition to the above scheme is to require each station to "listen before talking," i.e., require a station to sense the medium to determine if another station's signal is present and defer transmission if this is the case. This situation is shown in Figure 1.4.4 where a third station at the middle of the cable

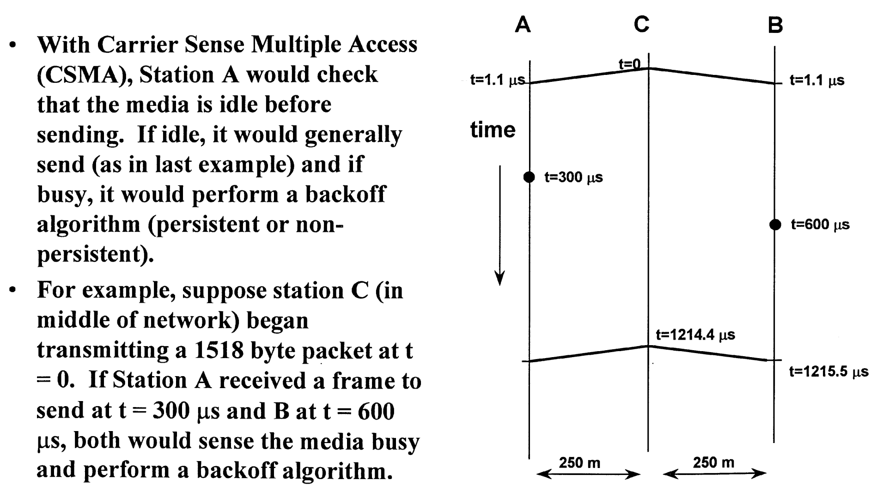

FIGURE 1.4.4 Use of carrier sense multiple access (CSMA).

TABLE 1.4.2 Typical Persistency Algorithms

Persistency Scheme	Description
Non-persistent	• idle ⇒ transmit • busy ⇒ wait random time and repeat
1-persistent	• idle ⇒ transmit • busy ⇒ wait until idle then transmit immediately (Note that if 2 or more stations are waiting to transmit, a collision is guaranteed)
p-persistent*	• idle ⇒ transmit with probability p and delay one time unit with probability 1-p; time unit is typically the maximum propagation delay • busy ⇒ continue to listen until channel is idle and repeat above for idle • delayed one time unit ⇒ repeat above for idle

* Issue is choice of p

- Need to avoid instability under heavy load.
- If n stations are waiting to send, the expected number transmitting is np. $np > 1 \Rightarrow$ collision is likely.
- New transmissions will also begin to compete with retries and network will collapse: all stations waiting to transmit, constant collisions, no throughput.
- Thus np must be <1; but heavy load means p must be small and time will be wasted even on a lightly loaded line, e.g., $p = 0.1 \Rightarrow$ on average, will transmit in tenth interval on an idle line.

begins sending at time t = 0. Due to signal propagation delays, signal reception begins at both A and B at time t = 1.1 μs. In the figure, while sensing the presence of carrier from C, A and B both receive frames from higher layers to transmit but, adhering to the CSMA scheme, defer transmitting until some time after station C's transmission is completed.

For typical CSMA schemes, a number of strategies can be employed to determine when to begin transmitting after deferring to a signal already on the medium. These strategies typically involve invoking one of the persistency schemes shown in Table 1.4.2. The persistency parameter "p" relates to the probability that a station sends its frame immediately after the medium is sensed idle. To obtain maximum channel utilization, the choice of the persistency value, 0.1, 0.2, 0.3, …, etc. is dependent on the traffic offered by the stations. A low level of traffic would operate best with a persistency value, p, near 1.0 (here, typically only a single station will be ready to send and thus should send immediately with high probability) and a high level of traffic would operate best with a lower value of p (here multiple stations will likely be ready to send and the lower value of p will make it more likely that only one station attempts to transmit). It should be noted that the above retransmission algorithm is not related to the binary exponential backoff algorithm discussed below, associated with resolving collisions. Also of note is that the IEEE 802.3 standard specifies the 1-persistent scheme, ensuring a collision if two or more stations are deferring to an ongoing transmission.

1.4.2.4 Adding Collision Detection

A problem with the CSMA scheme is depicted in Figure 1.4.5, where stations A and B both have something to send at t = 0 and, sensing the medium idle (no carrier) both begin transmission. For example, this case would occur if both have been deferring to another station transmitting on the medium and used the 1-persistent backoff scheme. At some short time later, the signals will collide at stations A and B (and at all other stations on the medium). In this case, no useful information is transferred for the entire transmission time of the frame, approximately 1200 μs for a frame of maximum length.

A solution to this problem is the addition of the collision detection mechanism depicted in Figure 1.4.6. The addition of such a mechanism reduces the wasted transmission time as both stations will stop transmitting upon detection of the collision. Here the stations have the added capability of detecting the occurrence of a "collision" of the two signals on the medium. With this added functionality, the stations can stop transmitting upon detecting collisions and immediately undertake a backoff scheme to allow one station to capture the medium.

- **Problem: Suppose both stations decide to transmit in the time frame t = 0 to t = 2.2 µs. With no collision detection, both will continue to transmit for the entire frame time.**
- **Considerable time will be wasted (over 1200 µs in this example).**
- **Also, a scheme is needed to determine who transmits when the media becomes idle or else, with probability 1, they will collide again, e.g., an n-persistent scheme (e.g., .1-persistent) could be used, i.e., a particular station transmits with probability n (e.g., probability of .1).**

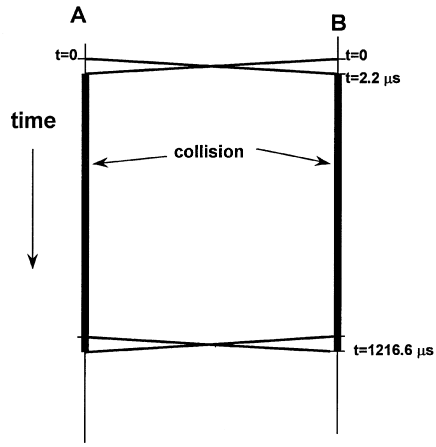

FIGURE 1.4.5 Wasted resources with CSMA.

- **Collision Detection (CD) solves this problem; the stations stop transmitting when they sense a collision. After they stop transmitting, they implement a *binary exponential backoff* scheme (random retransmission interval doubled each time a collision is detected).**
- **CSMA/CD also employs the 1-persistent scheme for stations accessing a media that just became idle. This can raise the chance of collision since multiple stations waiting for the media to become idle will collide but the collision detection scheme will minimize wasted resources.**

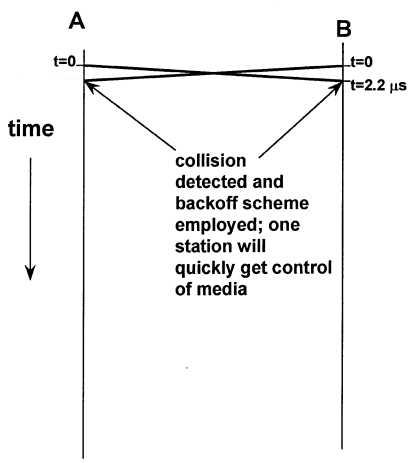

FIGURE 1.4.6 CSMA/CD.

1.4.2.4.1 Collision Backoff Scheme

Two stations, A and B, implementing the CSMA/CD media access control technique are shown in Figure 1.4.7. If, as shown in the figure, they both sense the media idle and begin to transmit at t = 0, a collision will occur and IEEE 802.3 specifies a "truncated binary exponential backoff" randomization scheme so that one of the stations can obtain control of the media. As shown in the figure, with this scheme, first one of two "slots" is chosen randomly by each station to attempt to capture the medium. The slot time is chosen based on factors that include the round trip transmission time between two stations at the ends of the medium, and the time required to detect a collision. It is specified in bit times; the IEEE 802.3 standard specifies a slot time of 512 bit times (51.2 µs for a 10 Mbps system). If a collision

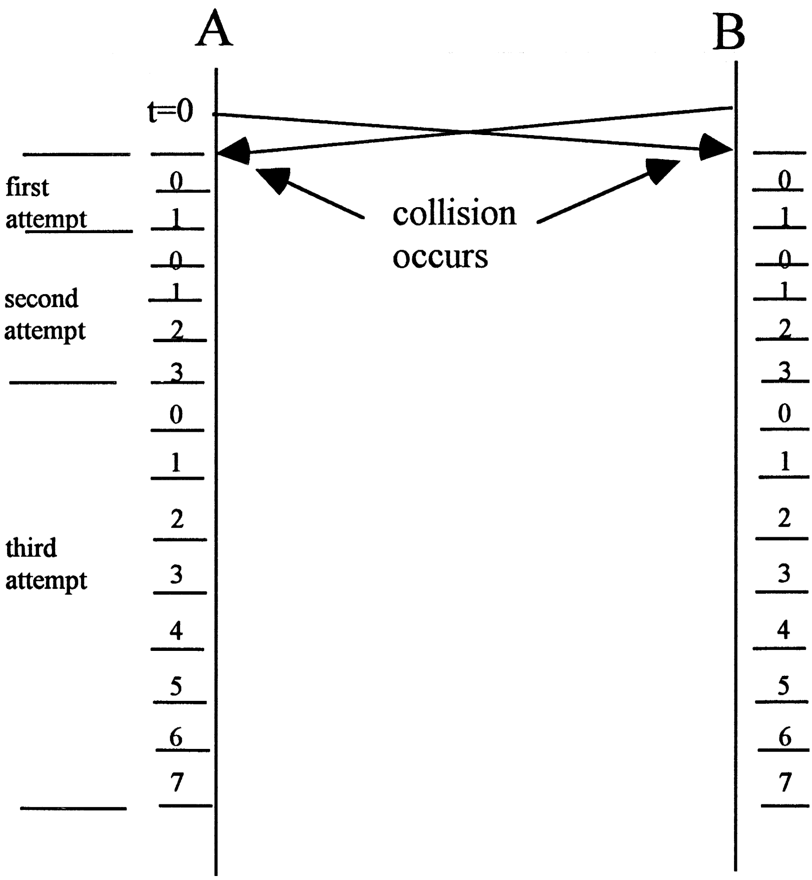

FIGURE 1.4.7 Retransmission attempts with the binary exponential backoff scheme.

occurs again (i.e., they both choose the same slot), one of four slots is chosen in the next attempt; then eight; then sixteen; etc. The number of slots grows in this manner to 2^{10} and truncates at this value. After a total of 16 retransmission attempts fail, this event is reported as an error.

With the CSMA/CD scheme, for reasonable traffic levels, a station should capture the medium in a rather short time, especially when compared to the CSMA scheme. For instance, consider the two stations in Figure 1.4.6 implementing a 1-persistent scheme (the recommended IEEE 802.3 scheme used after deferring to a transmitting station). Each will begin to transmit when station C stops transmitting and thus will suffer a collision on this first transmission. The binary exponential backoff scheme of Figure 1.4.7 will yield the following results.

For the two stations competing for the medium, the following outcomes are equally likely during the first retransmission.

(0,0), i.e., station A picks slot 0 and station B also picks slot 0,
(0,1), i.e., station A picks slot 0 and station B picks slot 1,
(1,0), i.e., station A picks slot 1 and station B picks slot 0,
(1,1), i.e., station A picks slot 1 and station B also picks slot 1.

Two of these outcomes, (0,1) and (1,0), will result in a station capturing the medium, station A in the first of these and station B in the latter. Two of the outcomes, (0,0) and (1,1) result in collisions in slots 0 and 1, respectively. Thus with a probability of _ = 0.5, one of the stations will capture the medium in this first retransmission period.

If there is a collision for the first retransmission, the second retransmission uses 4 slots chosen at random by the stations (numbered 0, 1, 2 and 3 in Figure 1.4.7) , resulting in 16 possible outcomes. Using similar notation as above, these outcomes are

(0,0), (0,1), (0,2), (0,3)
(1,0), (1,1), (1,2), (1,3)
(2,0), (2,1), (2,2), (2,3)
(3,0), (3,1), (3,2), (3,3)

and only four of these, (0,0), (1,1), (2,2), and (3,3), result in collisions yielding a probability of 12/16 or _ for a successful outcome. Thus the probability of exactly two retransmissions is (prob of collision on first retransmission) × (prob of success on second) = _ x _ = 3/8 = .375.

Similarly, the probability of exactly three retransmissions is

$$1/2 \times 1/4 \times 7/8 = 0.109,$$

and that for four is

$$1/2 \times 1/4 \times 1/8 \times 240/256 = 0.0146.$$

The likelihood of more than four transmissions is rather small.

Finally, the average number of retransmissions is

$$\sum_{i=1}^{\infty} i \times (\text{prob of i retransmission}) = (1 \times 0.5) + (2 \times 0.375) + (3 \times 0.109) + (4 \times 0.0146) \ldots$$

$$= 0.5 + 0.75 + 0.327 + 0.058 \ldots = 1.635$$

On average then, with two stations competing for the medium, one will capture the medium during the second retransmission attempt. Note that this saves medium resources when compared to Figure 1.4.5 where over 1200 μs are wasted due to the collision and additional time will be spent in some sort of backoff scheme.

It is interesting to note that for 3 stations competing, media capture will occur more quickly. Here, the three stations will collide on the first transmission attempt and the eight possible outcomes for the first retransmission are (0,0,0), (0,0,1), (0,1,0), (0,1,1), (1,0,0), (1,0,1), (1,1,0) and (1,1,1), with only (0,0,0) and (1,1,1) resulting in unsuccessful outcomes. For example, the (0,0,1) outcome will result in media capture: stations A and B will collide in slot 0 and station C will capture the medium in slot 1. Thus the probability of one station capturing the medium in the first retransmission is _, greater than the case with two stations competing. The average number of retransmissions for this case can be shown in the above manner to be on the order of 1.27. The reduced average number of retransmissions in the case of three stations competing is somewhat of an anomaly; for more stations competing, the average number of retransmissions steadily increases. Of course, the average waiting time for a particular station to capture the medium will increase with the number of stations competing for the medium.

1.4.2.5 CSMA/CD System Components

As mentioned, the IEEE 802.3 specifications support multiple media types, including coaxial cable, twisted pair, and fiber. Thus one of the component interfaces will of necessity vary with the media type. For example, the physical media dependent interface for twisted pair differs in a number of respects from that for fiber, including the physical connector (or plug), the electrical vs. optical nature of the interface

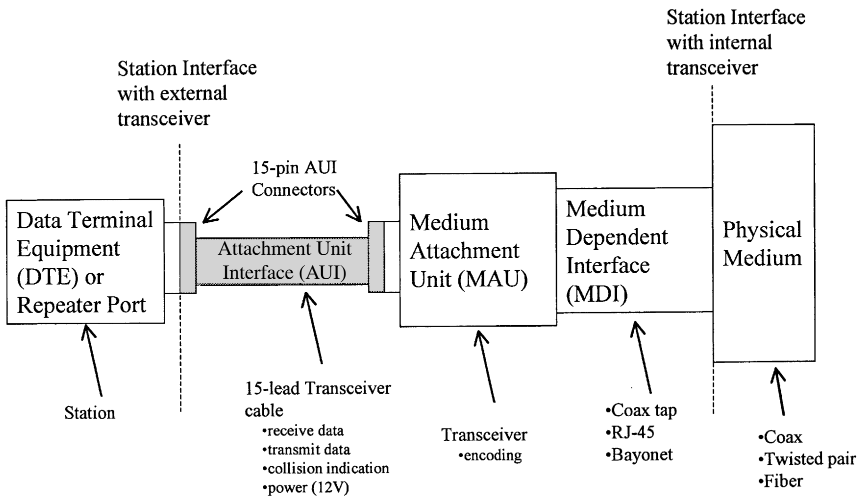

FIGURE 1.4.8 CSMA/CD system components.

and the encoding scheme (translating a logical sequence of bits to the electrical/optical signal on the medium), e.g., Manchester or Non-return to Zero (NRZ) schemes. In a number of IEEE implementations, this interface is remote from the station itself and the IEEE specification provides a media-independent manner of extending the station interface to the media-dependent interface.

This general situation is shown in Figure 1.4.8, where the various functional components that make up the CSMA/CD system are also shown. The Medium Dependent Interface (MDI) is shown on the right of the figure and provides the direct electrical/optical connection. Examples include an 8 pin connector (RJ-45 telephone style jack) for twisted pair, a coaxial cable clamp for coax medium and a spring-loaded bayonet connector (termed an ST connector) for fiber. The Medium Attachment Unit (MAU) shown in the figure, also commonly known as a transceiver, is also specific to the type of medium and performs signal encoding (e.g., Manchester or NRZ). Transceivers also contain a jabber protection circuit that protects the network from a station or transceiver that is transmitting frames whose length exceeds the maximum allowed.

The interconnection of the station to the remote transceiver is accomplished by the Attachment Unit Interface (AUI), also termed the transceiver cable. This 15-lead cable is media independent and carries data receive, data transmit and collision detection signals along with power from the station to the transceiver. The maximum length of the transceiver cable is 50 m. The associated 15 pin AIU connector is commonly found on Ethernet station cards.

It is important to note that many implementations include the transceiver as part of the Ethernet station card. For example, twisted pair implementations use this scheme. In this case, only the media-dependent interface in visible on the station Ethernet card; e.g., the RJ-45 type interface is commonly found on interface cards connecting to twisted pair medium.

1.4.2.6 Example Implementations

IEEE 802.3 standards are characterized by a shorthand notation to facilitate their description. The notation (e.g., 10BASE5) is composed of three elements as shown in Figure 1.4.9, indicating the transmission rate of the system in Mbps, the modulation scheme, baseband or broadband, and the maximum length of the segment in hundreds of meters. With standards adopted more recently, such as 10BASE-T, the IEEE has been more descriptive with its notation. For example, the "T" in the 10BASE-T notation is short for "twisted-pair wiring."

This section describes the most commonly implemented versions of the IEEE 802.3 specification.

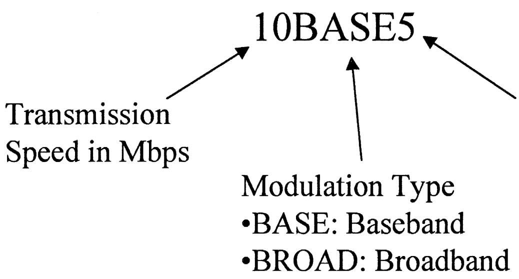

FIGURE 1.4.9 IEEE 802.3 nomenclature.

1.4.2.6.1 10BASE5

10BASE5 was the first version of the IEEE specification to be developed and it most closely resembles the earlier DIX versions 1 and 2. [4] The 10BASE5 specification employs a "thick Ethernet" 50-ohm coaxial cable. While this cable is difficult and relatively expensive to install, it provides advantages over other implementations in terms of distance and the number of terminations permitted for each segment.

This specification uses the standard outboard transceiver option discussed above: the station adapter board connects to the standard Attachment Unit Interface (AUI) cable; this in turn is connected to the transceiver which is connected to the Ethernet trunk cable via a "vampire" tap. Up to 100 devices can be placed on a 500-meter segment, with a maximum of 1024 devices on a multi-segment network, discussed below.

1.4.2.6.2 10BASE2

10BASE2 (also known as "thin Ethernet" or "Cheapernet") employs a thin flexible coaxial cable (RG-58). In earlier implementations, the transceiver functions were onboard the station and the connection to the station was by means of a media-dependent BNC "T" connector. To provide the flexibility to use the station board for either the 10BASE 5 or 10BASE2 systems, station boards have been developed which provide options for both external 10BASE5 "vampire" taps and 10BASE2 BNC connectors. Board manufacturers commonly provide boards with built-in transceivers that can be switched on or off for a particular application.

The standard 10BASE2 LAN can support only 30 terminations on each coaxial cable segment of 185 meters. While this may seem like a constraint, it is often adequate for most work area environments. Where a requirement exists for interconnecting multiple work areas, or work areas with multiple 10BASE2 segments, a backbone 10BASE5 segment can be employed to provide intersegment connectivity. Table 3 highlights the differences between the 10BASE5 and 10BASE2 systems.

1.4.2.6.3 1BASE5

This standard approach was contributed by AT&T to accommodate its earlier Starlan products. It operates at 1 Mbps, and as such is often most useful for small work areas or low-traffic environments. 1BASE5

TABLE 1.4.3 802.3 10BASE5/10BASE2 Comparison

	10BASE5	10BASE2
Common name	802.3 "Ethernet"	Cheapernet, THIN Ethernet, THINWIRE Ethernet, etc.
Type of cable	50 Ω Thick dual shield	50 Ω RG-58
Maximum segment length	500 m.	185 m.
Spacing of devices on cable	2.5 m. minimum	0.5 m. minimum
Maximum number of taps for a segment	100	30
Maximum number of full repeaters in a path between two stations	2	2
Type of taps	Vampire or coax	BNC "T" connector for "daisy chaining"

also employs inexpensive twisted-pair wire interconnected through a hierarchical system of concentrator hubs. The hubs emulate a bus configuration by broadcasting data and collision information on all ports.

1.4.2.6.4 10BASE-T

One of the most important developments in the IEEE 802.3 area was the specification of the 10 Mbps unshielded twisted-pair (UTP) Ethernet system, 10BASE-T. Virtually every vendor active in the Ethernet market now offers 10BASE-T products.

Like the 1BASE5 specification, this system uses a hub concentrator to interconnect multiple stations and emulate bus operation. These implementations are limited to 100-meter segments due to the greater attenuation and signaling difficulties of twisted pair. This does not present any unusual problems since these connections only reach to the communications closet. From there, fiber and coax segments can be used to concatenate and extend the LAN system. 10BASE-T systems use one twisted pair for transmitting data and a separate pair for receiving. "Collisions" are detected by sensing the simultaneous occurrence of a signal on both the transmit and receive pairs.

It is imperative, however, that organizations planning these networks have their existing twisted-pair wire certified for both attenuation and capacitance before making any assumptions on its reuseabilty.

1.4.2.6.5 10BASE-FL

The 10BASE-FL specification allows for the use of two fiber optic cables as the medium, one for signal transmission and one for reception. Such a medium allows advantages of greater distances, e.g., the standard allows segment lengths of up to 2000 meters, evolution to higher transmission speed, and isolation from electromagnetic radiation. The system components are identical to those shown in Figure 1.4.8 with the use of a fiber-optic Medium Attachment Unit (MAU).

1.4.2.6.6 10BROAD36

The 10BROAD36 implementation uses much of the same hardware as the baseband implementations. The specification enables an organization to use its existing workstation boards for connection to either a baseband or broadband system. The essential difference is the substitution of a broadband electronics unit and a passive broadband tap for the baseband MAU. The primary function of the broadband electronics unit is to create the frequency-derived data channels and to monitor for collisions. It also converts the signals from the baseband-coded signal on the AUI to the analog signal necessary on the broadband channel. Workstations can be placed up to 1800 meters from the "head-end" of the broadband cable plant. By placing the head-end in the center of the configuration, workstations can be installed up to 3600 meters from each other. In recent years, this standard has been less frequently used.

1.4.2.7 Topology Extensions

1.4.2.7.1 Repeaters

Repeaters regenerate the signals from one LAN segment for retransmission to all the others. The earliest repeaters were simple two-port devices that linked a couple of coaxial cable segments. Later, repeaters evolved to multiport devices deployed as the hub of a star topology. Since with repeaters, all segments are part of a unified LAN, the nature of the shared channel must be preserved by broadcasting all information to all attached devices. An aspect of these repeaters is that they must be capable of retransmitting collisions as well as data frames. In the case of IEEE 802.3 CSMA/CD LANs, physical LAN segment connection standards for repeaters are well developed and mature. The latest specifications for implementation of 10BASE-T repeaters are contained in the IEEE 802.3 specifications (see Section 9 of [8]).

In addition to the functions described above, repeaters can provide an optional "partitioning" feature between segments. This function is designed to address an abnormal situation such as a cable break or network card failure. Thus, if conditions on a given segment are causing an extensive proliferation of collisions, the rest of the LAN can be protected from this anomaly. The repeater will count the number of collisions from the source segment and when excessive, isolate these from transmission to the next segment.

1.4.2.7.2 Switched Hubs (Ethernet Switches)

In Ethernet switching, the interconnecting device (termed a switching hub) has intelligence to use the MAC-layer address of a received frame to determine the specific port on which the destination station is attached and transmit the frame on only that port. No other stations are aware of the frame. If frames arrive destined for a busy port, that port can momentarily holds them in its buffer; the size of port buffers differs by vendor from a few hundred to more than a thousand packets. When the busy port becomes free, frames are released from the buffer and sent to the port. This mechanism works well unless the buffer overflows, in which case packets are lost. To avoid this, some vendors offer a throttling capability; when a port's buffer begins to fill up, the hub begins to transmit packets back to the workstations. This effectively stops the stations from transmitting and relieves the congested state.

Some LAN switching products offer a choice of packet-switching modes, e.g., fast forward and "store and forward." These modes affect the amount of packet latency, the time from which the first byte of a packet is received until that byte is forwarded. Each mode reads a certain number of bytes of a packet before forwarding. This creates a trade-off among latency, collision, and error detection. The greater number of bytes read, the greater the latency, but the fewer errored or collision terminated frames propagated through the network. The fast-forward mode passes packets shortly after receiving the Destination Address portion of the Ethernet frame (see Figure 1.4.2) and, based on this field, determines the appropriate destination port. In this mode, typical latencies are on the order of tens of μs for a 10 Mbps Ethernet system. Store-and-forward mode receives entire frames and performs error detection via the FCS field (see Figure 1.4.2) before forwarding, resulting in increased latency but maximizing frame error detection. Here, for a maximum length frame, latency will be greater than 1200 μs. Some LAN switching devices support only 1 address per port, while others support 1500 or more. Some devices are capable of dynamically learning port addresses and allowing or disallowing new port addresses. Disallowing new port addresses enhances hub security; in ignoring new port addresses, the corresponding port is disabled, preventing unauthorized access. These and other techniques used in conjunction with port switching enhance overall network performance by eliminating the contention problem that occurs in shared Ethernet networks.

1.4.2.7.3 Multi-segment Guidelines

The IEEE 802.3 specifications provide guidelines for the number and types of segments that can be interconnected via repeaters and switch hubs. A number of example configurations are presented which would be typical of many implementations. For example, one possible configuration that taxes the CSMA/CD configuration guidelines has a distance of about 1800 meters between stations, with three 500 meter segments and two shorter segments interconnected by four repeaters. For more complex cases, the specifications provide guidelines for calculating system parameters such as the round-trip delay time and interframe gap shrinkage to ensure that these parameters are within allowed limits.

1.4.2.8 Higher Speed Extensions (100 Mbps and 1000 Mbps)

1.4.2.8.1 Fast Ethernet (100 Mbps) Overview

100 Mbps CSMA/CD operation, also termed Fast Ethernet, is specified in the IEEE 802.3u supplement to the standard. While attaining a ten-fold increase in transmission speed, other important aspects, including the frame format and the CSMA/CD access control scheme, remain unchanged from a 10 Mbps system, making the transition to higher speeds straightforward for network managers. The 100 Mbps specification uses Ethernet's traditional CSMA/CD protocol and is designed to work with existing medium types, including Category 3 and Category 5 twisted-pair and fiber media. In addition, Fast Ethernet will look identical to lower-speed Ethernet from the LLC layer upward. Since Fast Ethernet significantly leverages existing Ethernet technology, network managers will be able to use their existing knowledge base to manage and maintain Fast Ethernet networks.

The Fast Ethernet specifications include mechanisms for Auto-Negotiation of the media speed. This makes it possible to provide dual-speed Ethernet interfaces that can be installed and run automatically at either 10 Mbps or 100 Mbps.

There are three media varieties that have been specified for transmitting 100-Mbps Ethernet signals: 100BASE-T4, 100BASE-TX, and 100BASE-FX. The third part of the identifier provides an indication of the segment type.

"T4" is a twisted-pair segment that uses four pairs of telephone-grade twisted-pair wire.
"TX" segment type is a twisted-pair segment that uses two pairs of wires and is based on data-grade twisted-pair physical medium standards, covered later in this section.
"FX" segment type is a fiber-optic link segment that uses two strands of fiber cable and is based on the fiber optic physical medium standard developed by ANSI.

The TX and FX medium standards are collectively known as 100BASE-X. The 100BASE-TX and 100BASE-FX media standards used in Fast Ethernet are both adopted from physical media standards first developed by ANSI for the Fiber Distributed Data Interface (FDDI) LAN standard (ANSI standard X3T9.5), and are widely used in FDDI LANs. Rather than "re-inventing the wheel" when it came to signaling at 100 Mbps, the Fast Ethernet standard adapted these two ANSI media standards for use in the new Fast Ethernet medium specifications. The T4 standard was also provided to make it possible to use lower-quality twisted-pair wire for 100 Mbps Ethernet signals.

1.4.2.8.2 Gigabit Ethernet (1000 Mbps) Overview

Gigabit Ethernet, under the auspices of the IEEE 802.z working group, builds on the CSMA/CD MAC scheme and increases the transmission speed to 1000 Mbps. A key feature of Fast Ethernet implementations is the autoconfiguration capability, and Gigabit Ethernet solutions providing 10/100/1000 Mbps operation allow comparable features.

It should be noted that, as in the case for Fast Ethernet, a number of challenges involved in achieving rapid time to market for Gigabit Ethernet were resolved by merging existing technologies:

1. IEEE 802.3 CSMA/CD and
2. ANSI X3T11 Fibre Channel; Fibre Channel encoding/decoding integrated circuits (ICs) and optical components were readily available and optimized for high performance at relatively low cost.

Leveraging these two technologies meant that the Gigabit Ethernet standard could take advantage of the existing, proven high-speed physical interface technology of Fibre Channel while maintaining the IEEE 802.3 Ethernet frame format, backward compatibility for installed media, and use of CSMA/CD. This strategy helped minimize complexity and resulted in a technology that could be quickly standardized. Figure 1.4.10 shows how key components from each technology have been leveraged to form Gigabit Ethernet. As a result the Gigabit Ethernet standard based on fiber optics for the MAC and Physical layers has progressed rapidly. Unshielded twisted-pair media did not have the advantage of a proven, existing technology base and the standards remained in further development; this work is under the auspices of the IEEE 802.3 Working Group and referred to as 1000BASE-T.

Physical Layer Characteristics of Gigabit Ethernet

The initial Gigabit Ethernet specification from the IEEE 802.3z working group calls for three transmission media: single-mode and multimode fiber and balanced shielded 150-ohm copper cable. There are two supported types of multimode fiber: 62.5-micron and 50-micron diameter fibers. The IEEE 802.3ab committee is examining the use of unshielded twisted pair (UTP) cable for Gigabit Ethernet transmission (1000BASE-T). The distances for the media supported under the IEEE 802.3z standard and those projected for the IEEE 802.3ab are summarized in Figure 1.4.11.

Fiber Optic Media (1000Base-SX and 1000Base-LX)

As mentioned, the Fibre Channel physical medium dependent specification was employed for Gigabit Ethernet to speed standardization. This standard provides 1.062 gigabaud in full duplex mode and Gigabit Ethernet will increase this rate to 1.25 gigabaud with an 8B/10B encoding scheme allowing a data transmission rate of 1000 Mbps. In addition, the connector type for Fibre Channel was also specified for both single-mode and multimode fiber.

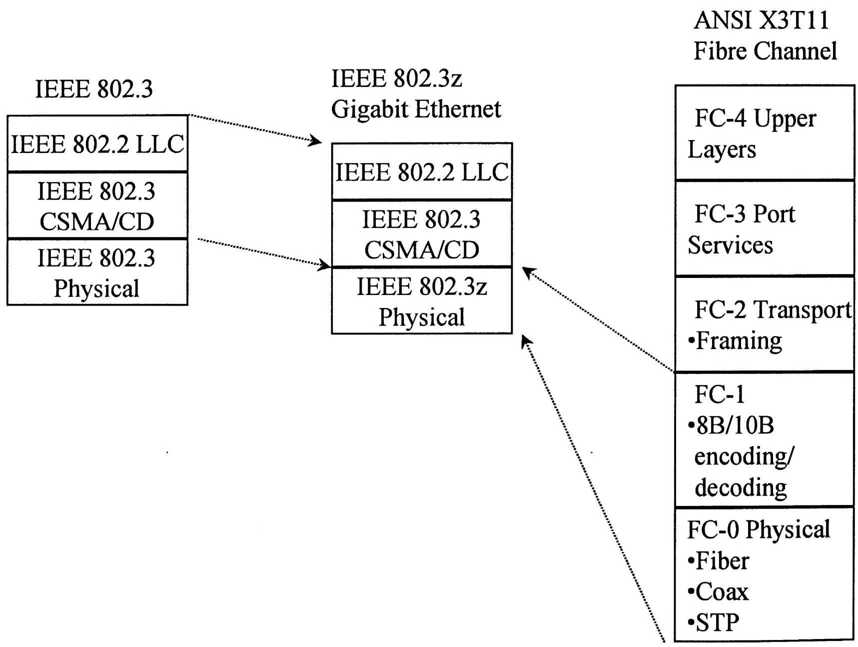

FIGURE 1.4.10 Gigabit Ethernet and the ANSI Fibre Channel Standard.

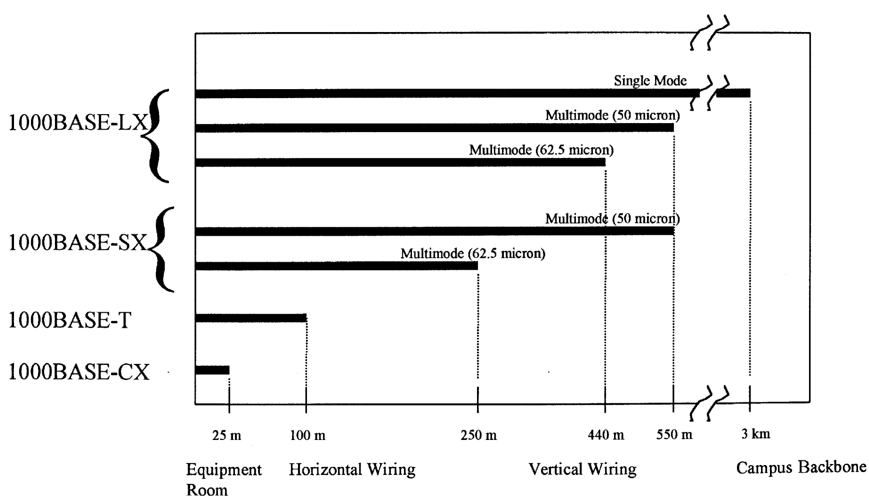

FIGURE 1.4.11 Distance specifications for gigabit ethernet media.

The standard supports two laser types, a short-wave laser type (termed 1000Base-SX) and a long-wave laser type (termed 1000Base-LX). Both short-wave and long-wave lasers are supported over multimode fiber. There is no support for short-wave laser over single-mode fiber. The key issues between the use of long-wave and short-wave laser technologies are cost and distance. Short-wave lasers are readily available since variations of these lasers are used in compact-disc technology. Long-wave lasers take advantage of attenuation dips at longer wavelengths in the cable and suffer lower attenuation. The net result is that short-wave lasers will cost less, but traverse a shorter distance. In contrast, long-wave lasers will be more expensive but will traverse longer distances.

The 62.5-micron fiber is typically seen in vertical campus and building cable plants and has been used for Ethernet, Fast Ethernet, and FDDI backbone traffic. However, this type of fiber has a lower modal bandwidth (the ability of the cable to transmit light), especially with short-wave lasers. This means that short-wave lasers over 62.5-micron will generally traverse shorter distances. The 50-micron fiber has significantly better modal bandwidth characteristics and will be able to traverse longer distances with short-wave lasers relative to 62.5-micron fiber.

150-Ohm Balanced Shielded Copper Cable (1000Base-CX)
For shorter cable runs (of 25 meters or less), Gigabit Ethernet will allow transmission over a new type of special balanced shielded 150-ohm cable (termed 1000Base-CX). Because of a distance limitation of 25 meters, this cable will likely have limited use.

1.4.2.9 Full Duplex Operation

The previous discussion has focused on half-duplex operation, where only a single communications channel was available and thus data could be transmitted in only one direction at a time. The CSMA/CD Media Access Control mechanism was required to determine which station could use the single channel. Full-duplex is an optional point-to-point mode of operation between a pair of devices allowing simultaneous communication between the devices and thus a doubling of aggregate capacity. Two separate communications channels are needed in this case to allow both stations to simultaneously transmit and receive. Thus the allowed physical media for this operation are only those with the capability of supporting two simultaneous channels, e.g., 10BASE-T provides independent transmit and receive data paths that can be simultaneously active. Full duplex operation cannot be supported on coaxial cable systems since they do not provide independent transmit and receive data paths. The optional full duplex mode of operation is specified by the 802.3x supplement to the standards.

1.4.3 IEEE 802.2 Logical Link Control Layer

The IEEE 802.2 Logical Link Control (LLC) layer specifications [9] include those Data Link Layer functions that are common to all 802 LAN MAC sublayer alternatives. The LLC frame format is shown in Figure 1.4.2.

Three basic types of service are defined in the standard.

Type 1 (Connectionless)
This service provides a best-effort delivery mechanism between origin and destination nodes. No call or logical circuit establishment procedures are invoked. Each frame is treated as an independent entity by the network. The type of frame used to provide this service is the unnumbered type; no flow control or acknowledgments are provided with this service. If the packet does not arrive at the destination, it is the responsibility of higher layers to resolve the problem through time-outs and retransmission. This type of service would be provided to the IP network layer protocol.

Type 2 (Connection Oriented)
Many wide area network protocols require that a logical circuit or call be established for the duration of the exchange between the origin and destination nodes. Packets usually travel in sequence over this logical circuit and are not routed as independent entities. LLC Type 2 provides this type of service. The service involves a number of control frames to manage the logical circuit (establishment, disconnection) and numbered frames for information transfer. Positive acknowledgments and flow control mechanisms based on this frame numbering are an integral part of this service. LLC Type 2 is commonly found in implementations of IBM's Systems Network Architecture (SNA).

Type 3 (Acknowledged Connectionless)
No circuit is established in this service variation, but acknowledgments are required from the destination node. This type of service adds additional reliability to Type 1, but without the overhead of Type 2.

These LLC alternatives are summarized in Table 1.4.4.

TABLE 1.4.4 Summary of Logical Link Control Alternatives

Service Type	Type 1	Type 2	Type 3
Description	Connectionless	Connection	Acknowledged Connectionless
Acknowledgments	No	Yes	Yes
Error recovery	No	Yes	Yes
Flow control	No	Yes	No

1.4.4 Building Cabling Specifications

The major components of a building cabling architecture that apply to the implementation of LANs are shown in Figure 1.4.12 and include the following.

- Equipment room — This location houses major data center processing and communications equipment for the building including servers, routers, and LAN switches. For a campus environment involving a number of buildings, one such location would serve as a data processing and communications center for the campus; other equipment rooms on the campus would serve specific buildings.
- Telecommunications closet — This is an area, typically located on each floor of a building, that houses data and telecommunications equipment providing wiring concentration, cross-connect and hubbing functions.
- Backbone cabling — This cabling provides connectivity between equipment in the equipment room and the telecommunications closets. It includes vertical connections between the floors and connections between buildings.
- Horizontal cabling — This cabling extends from the telecommunications closet to the individual work areas on the building floors.

The American National Standards Institute (ANSI), the Electronics Industry Association (EIA) and the Telecommunications Industry Association (TIA) develop specifications for commercial building

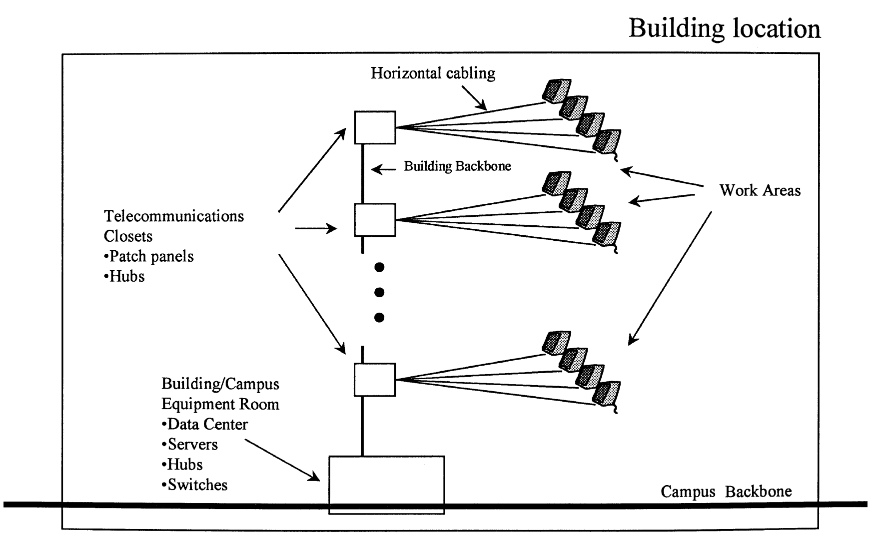

FIGURE 1.4.12 Architecture for building/campus telecommunications cabling.

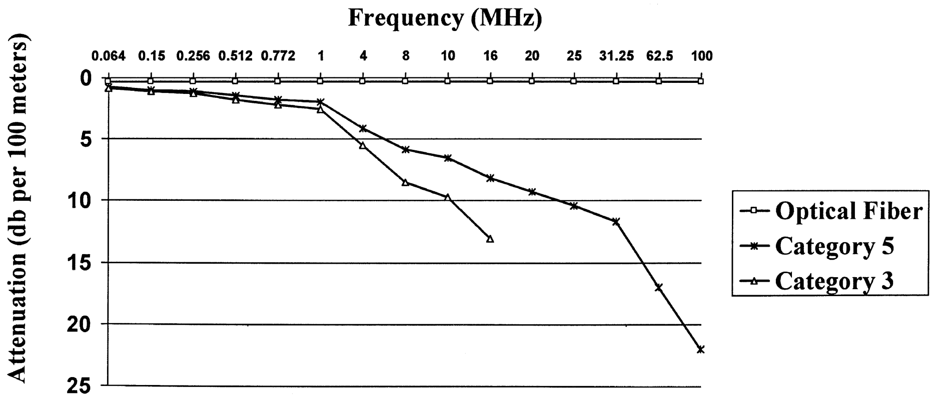

FIGURE 1.4.13 Attenuation of various media — EIA 568A.

cabling standards. This set of standards, referred to as ANSI/EIA/TIA 568A, defines the installation practices, certification, and physical, electrical, and optical characteristics for various physical media such as unshielded twisted-pair and fiber-optic cable [10]. The intent of the standard is to provide a guideline by which a cabling system can be designed and implemented as part of the overall design of a new building, even if the systems that the cabling must support are not yet defined. It guides the user toward the selection of cabling that will support current and future communications needs. The 568A standard and other EIA standards are recognized by architectural and engineering firms as definitive guidelines to use during the design phase of a building.

The 568A standard defines the requirements of a cabling system on a generic level that is appropriate to a commercial environment. The standard allows certain options, such as the use of various cabling media, including the following.

- 100-ohm unshielded twisted-pair cable in a four-pair configuration;
- 150-ohm shielded twisted-pair cable in a two-pair configuration;
- 50-ohm coaxial cable (not recommended for new installations); and
- 62.5 micron optical fiber cable in a two-pair configuration.

These cables exhibit performance that varies greatly depending on the frequency of the signal that is carried. For example, at Mbps speeds, a signal on a twisted-pair cable deteriorates in quality over a fairly short distance. The 568A standard provides performance criteria for the above cabling which must be met to be classified as 568A compliant. A summary of the attenuation limits specified by 568A for certain cable types is shown in Figure 1.4.13.

References

1. ISO/IEC TR 8802-1, Overview of LAN/MAN Standards.
2. IEEE Std. 802, Overview and Architecture.
3. ISO/IEC 7498-1: 1994, Open Systems Interconnection Basic Reference Model.
4. *The Ethernet, A Local Area Network, Data Link Layer and Physical Layer Specifications,* Digital Equipment Corp., Maynard, MA; Intel Corp., Santa Clara, CA; Xerox Corp., Stamford, CT; Version 1.0, Sept. 30, 1980, and Version 2.0, Nov. 1982.
5. IEEE 802.11, Wireless LAN Medium Access Control (MAC) Sublayer and Physical Layer Specifications.
6. IEEE 802.14, Standard Protocol for Cable-TV Based Broadband Communication Network.
7. RFC 1700, *ASSIGNED NUMBERS,* J. Reynolds, J. Postel. Oct. 1994.

8. International Standard ISO/IEC 8802-3: 1996(E), ANSI/IEEE Std 802.3, 1996 Edition, *Part 3: Carrier sense multiple access with collision detection (CSMA/CD) access method and physical layer specifications.*
9. ANSI/IEEE Std 802.2 [ISO/IEC 8802-3], Logical Link Control.
10. ANSI/EIA/TIA 568A, Commercial Building Telecommunications Cabling.

1.5 Token Ring Specifics

John Amoss

The IEEE token-passing ring local area network architecture, first approved in 1985 by the IEEE 802.5 Working Committee, has undergone a variety of additions and modifications and its essential specifications are complete. The token ring architecture specification addresses the media access control (MAC) layer and the physical layers shown in Figure 1.5.1 and the current version is defined in [1].

1.5.1 Topology

An IEEE 802.5 token ring consists of a star wired system of stations with each station connected by lobe cabling to a trunk coupling unit (TCU) on a concentrator. Figure 1.5.1 shows a sample token ring configuration consisting of two concentrators with three stations attached to each concentrator. The TCU provides a mechanism for insertion of a station into the ring and removal of the station from the ring. For example, note in the figure that station 5 is in bypass mode and is not participating in the ring operation. Concentrators can support multiple TCUs and are in turn serially connected via a trunk cable between ring in and ring out ports.

In addition to the architecture shown in the figure, a supplement to the IEEE 802.5 standard [2], defines a dual ring architecture intended for applications that require very high availability and recovery from media and station failures. This architecture uses two separate counter-rotating token passing rings; counter-rotating implies that information flow is in opposite directions on the two rings. Thus in addition to the ring shown in Figure 1.5.1, there is a separate independent secondary ring. The primary ring is normally the operational ring, with the secondary ring becoming operational in case of ring element

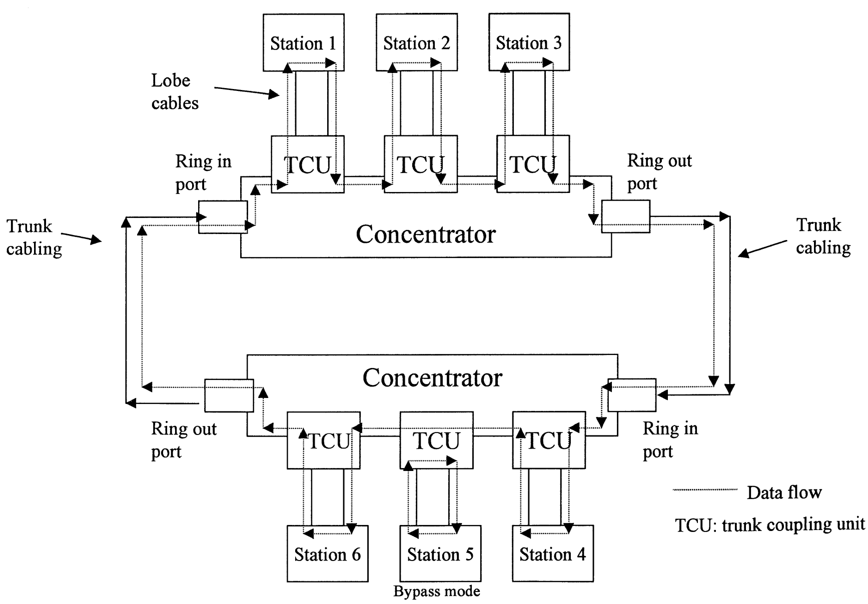

FIGURE 1.5.1 Examples of token passing rings confirmation.

failures. A typical reconfiguration to account for a failure condition involves wrapping the signal around from the primary ring to the secondary ring.

1.5.2 Station Attachment

As mentioned, each TCU provides insertion to or bypass from the ring for the station. Prior to insertion, while in the bypass mode, a station can perform self tests of the attaching lobe cabling and its station circuitry to assure proper operation. To attach to the ring, the station sends a signal over the lobe cable and the TCU switches from bypass to insert mode.

1.5.3 Token Ring Operation

Unlike CSMA/CD operation, each station regenerates and repeats each bit. Information on the token ring is transferred sequentially from one inserted station to the next. A given station transfers information onto the ring, where the information circulates from one station to the next. The addressed destination station copies the information and the station that transmitted the information removes it from the ring. Note also that the concentrators shown in Figure 1.5.1 may be passive (with no active elements) or active (performing a repeater function).

A station gains the right to transmit its information onto the medium when it detects a "token," a special media access control signal, passing on the medium. The token format is discussed in a later section. Any station may "capture" the token by modifying it slightly and appending additional fields including information to be transferred to a destination station. At the completion of the information transfer, the station initiates a new token, which provides other stations with access to the media.

The current specification includes an "Early Token Release" feature intended to make more efficient use of the available bandwidth on physically large rings operating with particularly small frames. In earlier versions of the token-passing protocol, a new free token could not be released by the sending station until it recognized the address in its own frame coming back around the ring to itself. If the frame was small, and the ring was large, there was a great deal of wasted time on the medium. Using Early Token Release, a sending station can release the free token immediately upon completing its transmission. The unused capacity on the ring can now be used by other stations. When coupled with the 16 Mbps ring operation, this new feature has significant advantages in terms of performance.

1.5.4 Priority Feature

An important feature of the IEEE token ring architecture, not found in the CSMA/CD architecture, is the support of multiple levels of priority. These priority levels are available for use by applications with varying classes of service needs. For example, the standard mentions real-time voice and network management as potentially high-priority applications. The priority mechanism operates in such a way that fairness is obtained for all stations within a priority level. This is accomplished by having the same station that raised the service priority level of the ring return the ring to the original service priority.

As mentioned, station access to the physical medium is controlled by passing a token around the ring. Operation of the priority feature is outlined in Figure 1.5.2 and is governed by the following parameters.

- P_{msg} = Priority of the frame to be transmitted by the station
- P_{rcvd} = Priority of the received token or frame (contained in the header)
- R_{rcvd} = Reservation level of the received token or frame (contained in the header)

As a token passes a station it has the opportunity to transmit one or more frames or place a request for a token of the appropriate priority. If a frame from another station passes the station, it may also place a request for a token of the appropriate priority. As shown in the figure, the station will send a frame if a token is received and the priority of the waiting frame is greater than or equal to the priority of the token. If the token has higher priority than the frame or if it is a frame from another station that is received, the station will attempt a reservation.

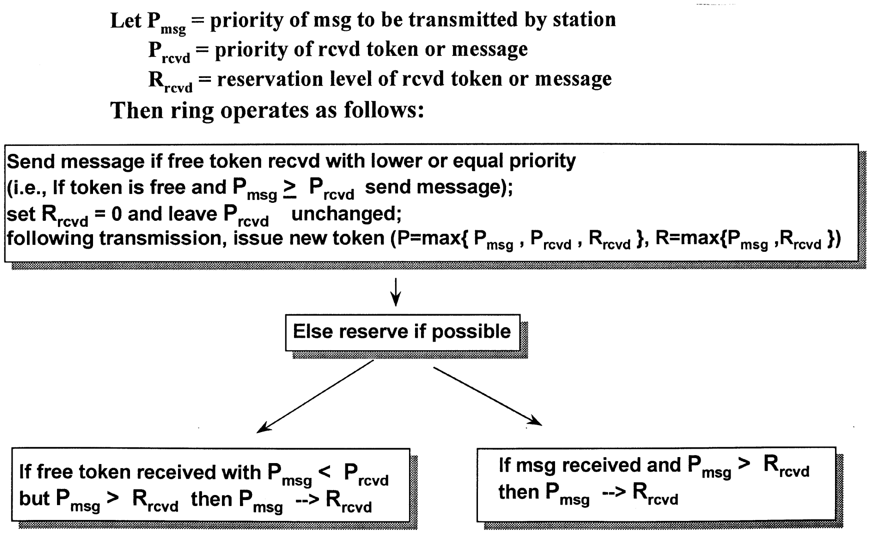

FIGURE 1.5.2 802.5 Token passing ring — priority and reservation scheme.

1.5.5 Management

Special stations are defined on the ring for system management purposes. Termed server stations, these stations act as data collection points on the ring, gathering information from the ring stations on any errors encountered, such as lost tokens, and lost or errored frames.

The server stations interact with a ring management system to provide the information necessary to manage the ring.

1.5.6 Physical Attributes

Distances covered by a token ring network are specified in terms of what is termed ring segment length, which is the transmission path between repeaters (i.e., stations or active concentrators). Since an active concentrator performs a repeater function, it can be seen from Figure 1.5.1 that allowable lobe cable lengths are longer for active than passive concentrators. For example, the token ring specification (Annex B) presents sample designs with lobe lengths on the order of 65 meters for passive concentrators and 200 meters for active concentrators.

The initial version of the IEEE token-passing ring was a 4 Mbps implementation which ran on shielded twisted-pair (STP) wire. The current token ring specification supports transmission rates of 4 and 16 Mbps and both STP and unshielded twisted pair (UTP) media.

Based on a number of factors such as the media type, data rate, and concentrator type, the number of station on a ring may be up to 250.

1.5.7 Formats

1.5.7.1 Tokens

The token, or control signal, has the format shown in Figure 1.5.3. This signal is the means of passing the right to transmit from station to station on the ring. As shown in the figure, the token consists of three octets.

The starting delimiter (SD) and ending delimiter (ED) each consist of a fixed sequence of symbols, some of which deliberately violate the differential Manchester encoding scheme (see Figure 1.5.4). Because of these violations, these delimiters cannot be part of the fill sequence, defined as a stream of validly encoded symbols, sent on the media between frames. Thus false SD and ED indications will not occur.

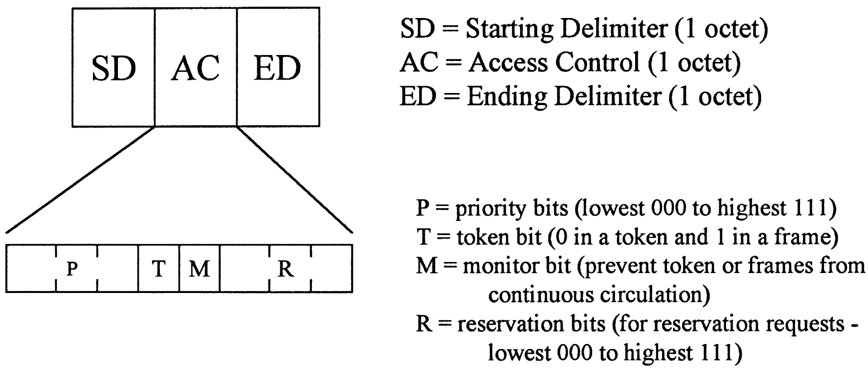

FIGURE 1.5.3 Token format.

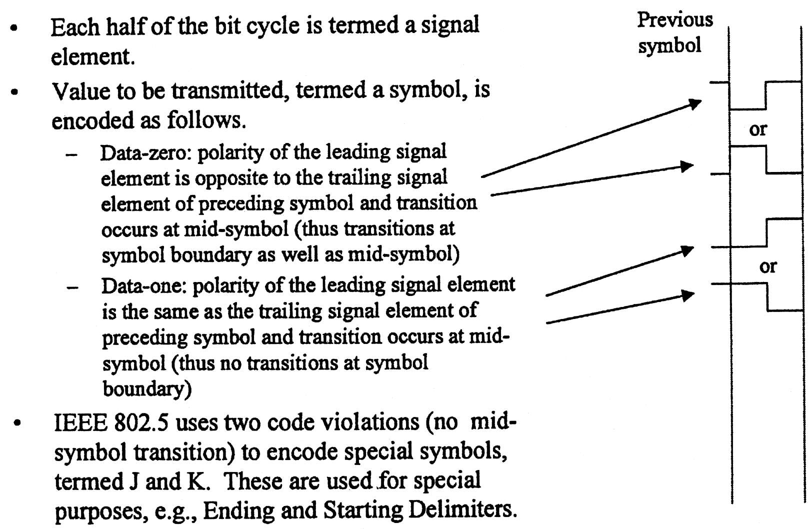

FIGURE 1.5.4 Differential Manchester encoding.

The access control field is a key field associated with many of the services provided by a token ring network.

- A 3-bit priority field indicates the priority of the token. When frames are eventually sent on the media, they carry the same priority as the token by which they gained access to the network.
- A 1-bit field indicates whether this is a token (value of 0), or the beginning of a frame transmitted by a station (value of 1).
- A 1-bit monitor field, set initially to 0 and set to 1 by a monitoring station, is used to prevent certain tokens or data frames from continuously circulating around the ring.
- A 3-bit reservation field allows stations with higher priority protocol data units to gain quicker access to the media.

1.5.7.2 Frames

Frames are the resulting messages sent between stations. It is important to note that two very different types of frames may be sent. One carries LLC messages, or user data, containing higher level protocol

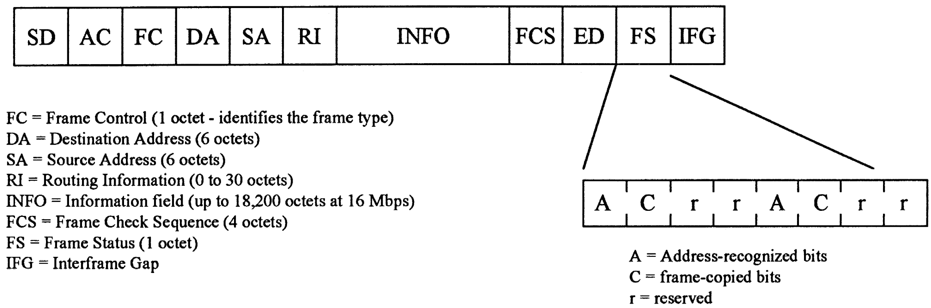

FIGURE 1.5.5 Frame format.

exchanges as described in the MAC section. A second type of frame, described in detail in the IEEE 802.5 standard carries MAC messages used to manage the MAC layer.

The format of these frames is shown in Figure 1.5.5.

In addition to those fields contained in a token, these frames contain additional fields, including the following.

- A Frame Control (FC) field indicates the frame type – MAC (either a MAC protocol frame or a MAC management frame) or LLC.
- A 48-bit source address (SA) field and a 48-bit destination address (DA) field identify the sending station and the receiving station. The destination address can represent an individual station, a group of stations or all stations (broadcast). The source address field contains 1 bit which is used to indicate whether routing information is included in the frame, as would be the case for a source-routed network.
- A routing information (RI) field is used in the case of source routing to specify the route for the frame through a bridged network.
- The information field contains octets destined for the MAC Protocol or Management entity or the LLC entity. A 32-bit frame check sequence field is used for error detection.
- A frame status (FS) field is used to provide an indication as to whether or not a frame reached its destination and whether it was successfully read from the media. "A" bits are set by a destination station to indicate that it has recognized its address in the DA field and "C" bits are set to indicate that it has successfully copied the frame.
- Finally, an interframe gap (IFG) field is added to account for variability in the gap between frames.

References

1. *Token ring access method and physical layer specifications*, ISO/IEC 8802-5 : 1998, ANSI/IEEE Standard 802.5, 1998 Edition.
2. *Recommended practice for dual ring operation with wrapback reconfiguration,* IEEE Standard 802.5c-1991, Supplement to 802.5

1.6 Summary

This section presented the current trends in voice and data networks regarding voice services; the evolution from the connection oriented networks to the connectionless networks, and unification between voice and data networks.

An Intelligent Network (IN) is a telephone network with a "service specific" architecture. It evolved in an Advanced Intelligent Network (AIN), which has a "service independent" architecture. This means that a given part of a phone number won't trigger a request for a specific service, but it can be interpreted differently by various services, depending on different factors such as time of day, caller identity, and type of call.

CTI is a technology platform that merges voice and data services at the functional level to add tangible benefits to business applications. In fact, CTI is a new application for pre-existing technologies. The most important benefit CTI brings to the corporate world is the potential to reduce operating expenses.

The low cost of the public Internet resulted in a great increase of data traffic in the last couple of years. Besides this, the quality of service (QoS) has been considerably improved, and offered acceptable delays for traffic less tolerant to delays. This raised a considerable interest in transporting packetized voice (not tolerant to delays) over data, in particular, IP networks. The support for voice communications over the Internet Protocol is usually called Voice over IP (VoIP).

Local area networks don't lose their importance for users independently of how much bandwidth is available in the wide area. They serve as local connectors between end-user-devices, server farms, storage farms, and the ingress/engress nodes of wide area networks. Their limitation to data is changing by supporting multimedia applications up to the desktop. Step-by-step, throughput capabilities could be improved; in this respect, Giga Ethernet seems to be the winner for high-speed LAN technology.