3 Network communications

OVERVIEW

Network communications provide the infrastructure to connect GIS operations. Networks connect user applications with shared data resources, remote office workflows with corporate data centers, and enable GIS users throughout the community and nation to share GIS data and services. Many geographic information systems today are globally connected, offering real-time information products to serve users around the world.

In providing an interactive model of our world, GIS uses a variety of data types (such as satellite imagery and aerial photography) to identify points, polygons, and lines that can be displayed on maps to represent spatial relationships. Often, these various data types must be brought together from a variety of data sources distributed in multiple locations. Nonetheless, the map display is rendered within seconds, with hundreds of spatial features converging. The GIS user may view the display for a few seconds, only to request a new map display (a different place, another resolution, etc.) again and again. All of this real-time (dynamic) map production—back and forth between the data source and the client application—can generate a considerable amount of traffic over the network. Therefore, understanding how much traffic your workflows will generate at peak times—and how much the network will bear—becomes a critical part of the system architecture design process.

Everyone can benefit from a fundamental understanding of network communications, but for decision makers that benefit resides on the bottom line. Although this may change in the future, the electronic products that compose this communications infrastructure can represent a major portion of the system budget. For that sort of investment, you want to be sure the network meets organizational needs, and that requires understanding an organization's workflows. Performance and scalability are the two signs of a computer system's efficiency, but both can be hamstrung by an existing network that can't stand up to how much you're asking it to do. GIS introduces a lot more data to move. Yet the answer is not as simple as "bigger is better." The bigger the network bandwidth, the higher the cost; and why pay for size you don't yet need? The answer lies in knowing your peak traffic loads, the network bandwidth that can handle those loads, and doubling the bandwidth to ensure optimum traffic flow.

While serving as an introduction to network technology in general, this chapter identifies the network protocols used with GIS specifically. These include standard IP disk-mounting protocols for accessing file data sources, message protocols for communicating between the GIS application and database data sources, and Web protocols for connecting intranet and Internet data sources and services. Design standards are provided for each supported communication protocol, along with general design guidelines for network bandwidth sizing. You will also find some standard network design planning factors. These factors, expressed in terms of traffic per map display, will be the numbers used for capacity planning purposes in later chapters.

The fundamentals

For GIS, a network is like a transportation system for data, and there are two classes of the technology, the first less expensive than the second: local area network (LAN) and wide area network (WAN). (Actually, the cost for LAN equipment is relatively inexpensive in general, compared to other hardware costs within the system environment.) A LAN can handle heavy traffic over a short distance, as on a college campus or in a building. The Web is an example of a WAN, which supports communication or data transport over long distances. An organization's WAN can also become a way of sharing data between various departments as well as a way to leverage service-oriented Web operations through the latest integrative technology (i.e., ArcGIS Server 9.3).

For both a LAN and WAN, the volume of data (measured in bits) that can be transported per second is referred to as the capacity or transport rate of a particular network segment. This capacity is called *network bandwidth* and is typically measured in millions of bits (megabits or Mb) or billions of bits (gigabits or Gb) per second. Bandwidth specifications (the rate of moving data) provide a simple way for GIS design architects to calculate and talk about an organization's network capacity needs.

Those needs are best expressed in terms of what bandwidth is necessary to handle the network traffic during the hours when the most work is done. If you ever doubt the importance of considering network traffic in your design analysis, remember the analogy of an old transportation system that hasn't been upgraded to accommodate population growth: too many cars, coming from all directions, funneling down to one lane. During peak work periods, operational workflow performance can slow to a crawl similar to what is experienced driving onto major highway arteries during rush hour—it seems as if it takes forever to get anywhere or you just can't get anywhere at all. In fact, sometimes it is the latter and networks do "crash," temporarily refusing to function at all. Insufficient bandwidth capacity is at the root of many remote client performance problems. Data has to move at a rate appropriate to supporting user productivity during peak business hours. Therefore, calculating what will be a sufficient bandwidth is a critical part of the network component of system design.

To help you do that, best practices in terms of GIS network traffic transport times are identified later in this chapter (figure 3-5) for each of the configuration alternatives currently available for client/server communications. You will see in the chart that desktop client/server applications perform best in a LAN environment. Remote desktop users should be supported from a Windows Terminal Server or distributed desktop environment: for example, the application runs in the computer room with local data access, and persistent remote client access is provided for active session display and control. Web services, on the other hand, work fine over widely distributed WAN and Internet environments.

Eventually, you can use the capacity planning methodology described in this book to identify the bandwidth your GIS operations require. But clearly, capacity is not the only factor to consider in planning network communications. The network's configuration must also be a fit; it, too, influences how fast data appears on your screen. Where are your data sources in relation to the desktop users? How much data will your applications require from one source as opposed to another? Also, the appropriate conventions (protocols) must be in place, to allow the network to interface with the various computer products and multiple data formats encountered among its members.

Smaller factors within bandwidth, configuration, and protocols play a role in data transport and therefore in how fast and reliably data traffic moves through your network. In turn, how well the network performs is a major factor in overall system performance. Cumulatively, every factor down to the smallest either detracts from or contributes to creating and maintaining optimal system performance. Right now, during initial system design, you simply need to size and configure a network that will support the system during implementation. You can test and tune it later to verify your applications perform as expected. Nonetheless, understanding network components and processes and their interrelationships well enough to model them is prerequisite to creating a system that works at all, and this chapter will help you in this. Math and physics provide wonderful tools to model our world down to the smallest interaction. Over the years, we have used these tools to model the world of network communications. As a result, we can offer network design planning guidelines and best practices in this chapter that can help you get it right. But first, you need to understand that world: the physical components and processes of a network communication infrastructure for GIS. Then you can model the interactions between them (using the capacity planning models and workflow performance targets introduced in this book) according to the realities of your situation. Doing so can get you within range of a network bandwidth and configuration adequate for your needs. And that's where you need to start—with sufficient bandwidth to support your peak traffic needs.

Network components and GIS operations

GIS is different than other information systems, and its data traffic may put a bigger load on the network than it's accustomed to. GIS operations rate among the heaviest data movers (joined by document management and video conferencing enterprise solutions). Geographical information systems are data heavy because geography is data rich, and the beauty of GIS analysis is in its ability to examine high volumes of data quickly and turn it into information that is useful to you (an information product). What used to take hours and days of research and analysis, GIS can do in moments. With information layered so that you can readily see its relationship to place (and time) and other information, GIS information products are typically represented in a user-friendly map display. What goes on behind the scenes to create this user-friendly experience—transforming complexity into simplicity—is the stuff of system design.

LANs and WANs

The fundamentals of network technology set the scene. For many years (1970s–1990s), network technology remained a relatively static environment, while computer performance increased at an accelerating rate. Recent advances in communication technology, however, have enabled a dramatic shift in network solutions—and your system design options. Worldwide communications over the Internet bring information from millions of sources directly to the desktop in real time. Wireless communications is fast becoming mainstream; now data can virtually be transmitted from any place at any time.

There are also a variety of physical media used to transport data, some of which may be among your network segments. Data is typically transported from one server to another over physical networks made of copper wire or glass fiber (LANs and WANs). Other types of transport media include microwave, radio wave, and satellite digital transmissions. Wireless radio frequency bands and laser beams are also used as a communication medium.

Normally, it's the transmission medium that limits how fast the data can be transmitted. This rate of data transport is identified by the specifications applied to its communication procedure, called the network *protocol*. Data and applications may reside at many sites throughout an organization. Allowing for the fact that data may also exist in different formats at these various sites, a protocol is a set of conventions that governs how the network treats the data. For example, protocols may specify different levels of compression that can reduce the volume of data, thereby increasing transport efficiency. Today, network products are made to encourage a stable and dependable environment overall for data transport, whatever communication methods are used among the variety of protocols that allow applications and data resources to be useful and shared wherever they are located.

Network transport solutions can be grouped into two general technology classes. Figure 3-1 illustrates these two types of networks and some of the fundamental terms associated with the technology of each.

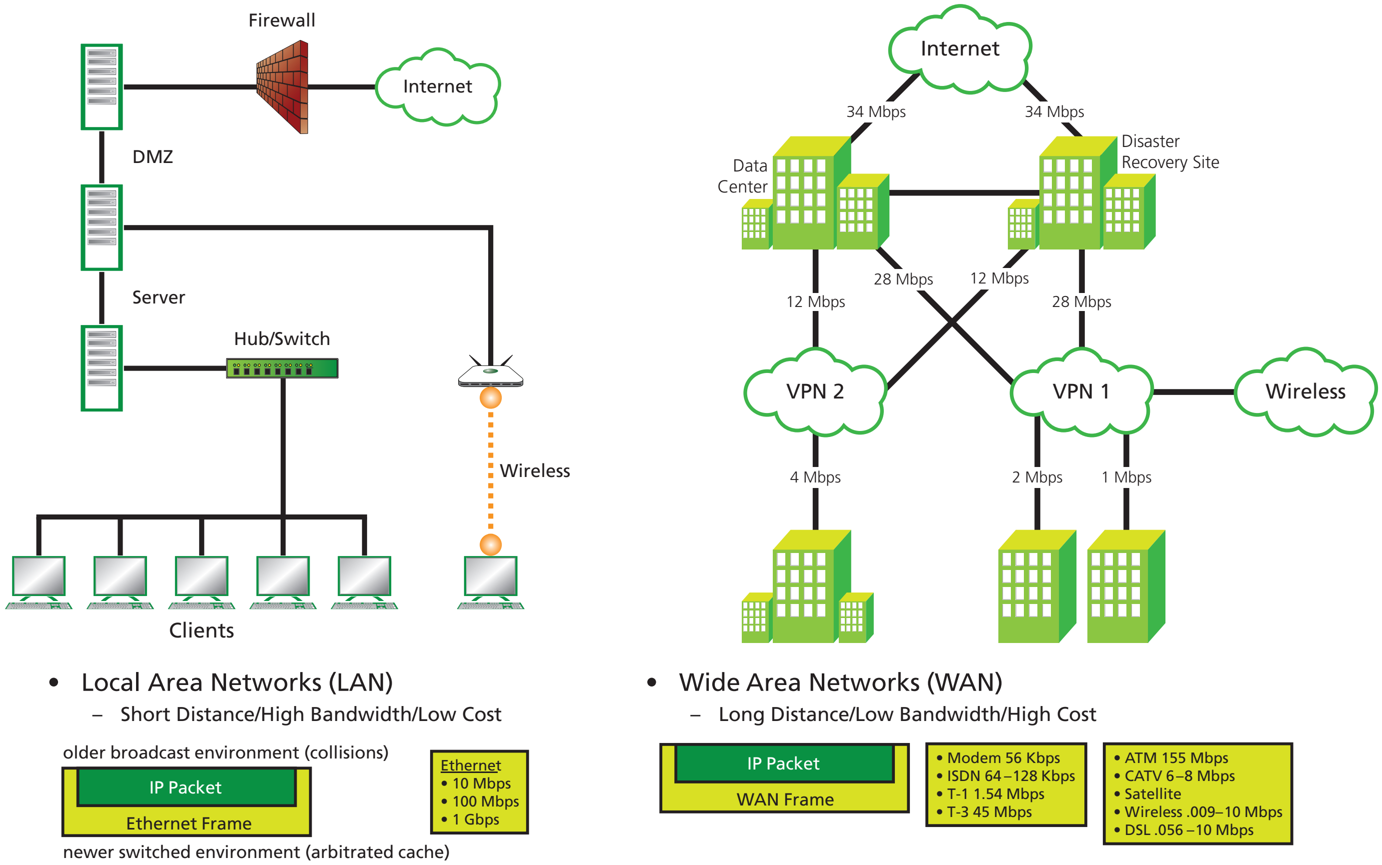


Figure 3-1

Types of networks.

Local area networks (LAN)

A LAN supports high-bandwidth communication over short distances. Data transport over a single technology like this is single threaded, which means only one data transmission can be supported on a single LAN segment at any time.

The local operating environment is much more efficient now than it was when local area network technology began. LAN protocols were initially introduced in the 1970s to provide electronic communications between central compute-server environments. By the early 1990s, several LAN protocols were in use, and protocol exchange devices (bridges) were needed to connect the different LAN environments (Microsoft Windows operating system, AppleTalk network system, UNIX operating system, Novell's network software, etc.). By the late 1990s, Ethernet technology became the LAN standard, and today Ethernet bandwidths provide the capacity to transport volumes of data per second; measured in megabits and gigabits as 10 Mbps, 100 Mbps, 1 Gbps and 10 Ggps. The bigger bandwidths are used in computer rooms and between buildings over campus-type environments. User desktops are connected by dedicated network cable to a central wiring closet where the network switches are located. Higher-bandwidth shared connections are provided between closets to the central computer facility.

Wide area networks (WAN)

A WAN supports long distance communications between remote locations. WAN protocols typically provide much lower bandwidth than what is available for the LAN, but the WAN is a data transport environment nonetheless: the source data is packaged in a series of packets and transported as a stream of packages (data packets) along the transmission medium. As mentioned earlier, the cost for WAN connections is relatively high compared to LAN environments.

The number of WAN protocols has expanded rapidly since the introduction of the Internet, to allow for the expanding communication medium, including telephone lines, cable TV lines, satellite communications, infrared transmitters, and wireless radio frequency bands, to name just a few. Data is transferred by light, electricity, radio frequencies, and laser beams—and these are only the common examples we see today.

Data

As you either know or have surmised by now, in GIS terms and in computer terms in general, data is essentially a collection of digital computer information stored in media that have the capability to record and retain the data structure. This data is represented by little pieces of information called *bits*. Each bit takes up the same space on storage or transmission medium. For convenience, these little bits can be grouped into *bytes* of data, with each byte containing eight bits. Data can be transported from one location to another within packets that protect the integrity of the data.

GIS users identify data capacity in terms of megabytes or gigabytes in reference to data stored on a computer disk. Network administrators, on the other hand, tend to think of data in terms of bandwidth specifications, identifying data in megabits or gigabits as the peak traffic volume a network can support per second. Megabyte is abbreviated using a capital "B," while megabit is abbreviated using a lowercase "b." These subtle differences can cause confusion when GIS specialists and systems administrators work together to design a system for GIS. So remember, 1MB = 8Mb when converting data volume from disk storage to data as traffic, and be very sensitive about using the proper abbreviation.

Another thing to remember: Network traffic includes some protocol overhead. The amount of overhead depends on the protocol packaging (size of network packets and the data volume). A simple guideline: translate 1MB of data to about 10 Mb of traffic.

What is data really?

The dictionary defines how we use the term in our language, and in today's world *data* is most commonly associated with computer technology and science. This tells us *how we use* the word, but does not tell us *what it is*. In a computer, it is represented in terms of switches (as a pattern of ones and zeros). In a communication medium, data is represented in frequency patterns, transmission patterns, etc. Data really is a pattern that represents a thought, an idea, or something we see or imagine (for example, a picture). That's not the whole story, but we're getting closer.

Where did it come from?

One of the first recorded datasets represented the names of animals or plants. We find pictures in ancient caves that tell us stories. Computers exist because someone thought of a way to represent our thoughts and observations by patterns that could be manipulated by computer processors and stored on disk (on and off switches). Yet, all this and we still can find no way to define data except in terms of representations that have to be used or processed to make them meaningful.

Does it actually exist?

The physicist's answer is no, data does not exist in the usual way we think of as "being." In itself—stripped of all means of conveyance—it is definitely not something we can touch, smell, or lift. We can observe how it is depicted, and we can translate this impostor from one medium into another. But there is no physical requirement for data itself to comply with the laws of physics—it has no weight or mass that must be moved when transmitted from one location to another. The limitations exist only by way of the medium in which data is represented. In other words, in system design, the medium helps us understand the message.

No substance, no weight, no mass, just a pattern representing a thought. This suggests there may be no physical limitations in how we move data—nothing really to limit data communication or what we think of as bandwidth capacity anyway. So then, might we not expect current infrastructure limitations to be resolved, over time, as we find more efficient ways to record (store) and communicate (transport) our ideas and observations (data)? It's just a thought. And one not half as unlikely as "Beam me up, Scotty," although it does have something in common with that Star Trek phenomenon: The only requirement (*and this is not small*) is that the pattern that arrives at the client processor is exactly the same as the pattern sent by the server processor.

Communication protocols

For arriving in the same shape it was sent, data owes its thanks to protocols, (or procedures) that make sure it does. Applications move data over the network through proprietary client/server communication protocols, which work this way: Communication processes located on the client and server platforms define the communication format and address information. Data is packaged in communication packets, which contain the communication control information required to transport data from its source client process to the destination server process while maintaining the data's original structure.

Communication packet structure

So many bits, so little time; the possibility for corruption of the data while on the road would be great were it not for the packet structure in which it travels. Data

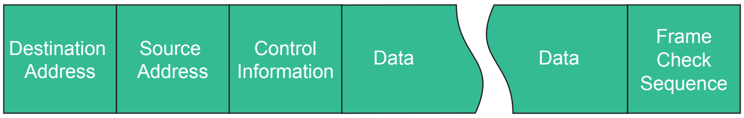


Figure 3-2

Communication packet structure for data delivery across a network.

is transmitted within packets that protect the integrity of the data. Within these packets is information that allows for its delivery across the network medium. The basic Internet protocol (IP) packet structure, as shown in figure 3-2, includes destination and source addresses and a series of control information, in addition to the data structure itself. Multiple packets are used to support a single data transfer.

Network transport protocol

The framework for client-server communications over a network—let's say, host-to-host over the Internet—is a series of step-by-step processes, with each protocol like a gateway to the next step. The terms "communication packet," "packet structure," and "data frame" are often used interchangeably; we'll simply use "packet" here, because "data frame" means something else in ESRI software terminology. But in describing how network transmission operates, let's be precise: the communication that goes across the network from host to host includes everything the packet structure is composed of (which has to be built every time). Just as we think of data in terms of the medium that moves it, we can think of that medium—in this case, the packet—in terms of the process that constructs it.

The packet is constructed at different layers during the transmission process. Data starts out in a stream to become a framework acceptable for travel across the network only after going through a layered experience. Using the standard way network administrators view the protocol stack (application layer, transport layer, Internet layer, and network access layer), figure 3-3 shows how a data stream from the host A application is sent through the protocol layers to establish a packet (data frame) with access to network transmission:

- The transmission control protocol (TCP) header packages the data at the transport layer.
- The Internet protocol (IP) header is added at the Internet layer.
- The medium access control (MAC) address information is included at the physical network layer.

The packet is then transmitted across the network to the host B side of the pendulum where the process—in reverse—moves the data to the host B application. A single data transfer can include several communications back and forth between the host applications, each time through these sequences.

GIS communication protocols

Figure 3-4 on the opposite page shows the primary communication protocols that GIS applications use for network data transfer. Component processes of client and server both take part in implementing each protocol, enabling delivery in this way: the client process prepares the data for transmission, and the server process delivers the data to the application environment where analysis and display take place.

Network file services and common Internet file system protocols

All GIS applications are able to access a variety of file formats from a local disk. Shared data can be provided over the network on a file share. The server platform operating system includes a remote disk-mounting protocol enabling the client application to access data from a distributed server platform. UNIX and Windows each

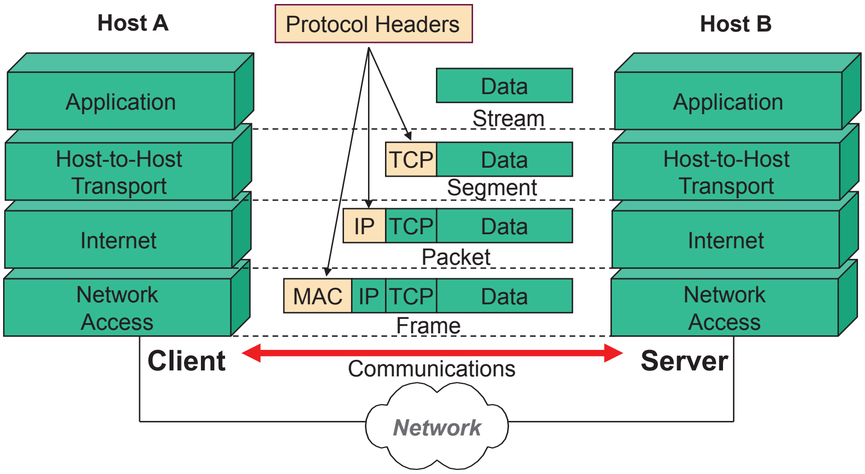


Figure 3-3

Network transport protocol.

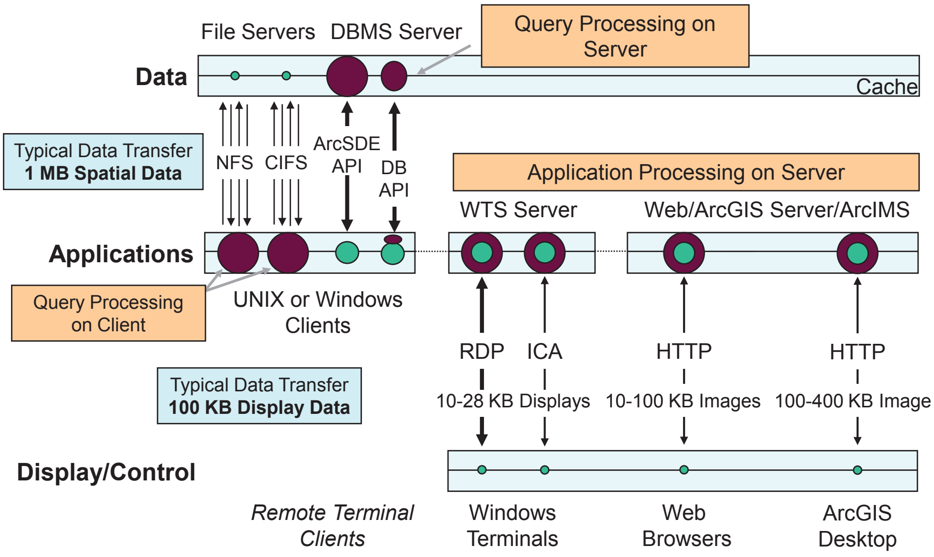


Figure 3-4

GIS communication protocols.

offer their own network-mounting protocols. UNIX provides a network file services (NFS) protocol and Windows provides a common Internet file system (CIFS) protocol, each allowing the client application to access the remote file share as if it were a local drive. The client and server platform must be configured with the same protocol stack to support remote file access.

When accessing data from a file server, all program-specific executables reside on the client platform. They provide direction to the server operating system, through the connection protocol for access to data located on the server platform. Each piece of data must be transferred to the client application to enable query, analysis, and display.

Many GIS and image file formats are optimized to minimize the volume of data that has to be transferred. The file (which might be quite large) can include an index, which the client application can use to identify the specific segment of the data file required to respond to the query. The client can then request solely the portion of the file needed for the display. These types of data structures improve input/output (I/O) performance when accessing large files.

Much more data must be managed in client memory when working with the file data source as a whole, as opposed to a selected segment of the file, so client memory requirements are much higher when accessing file data sources.

Database access protocols

ArcSDE (Spatial Database Engine) technology provides a schema and communication protocol that enables spatial data to be centrally managed within a commercial database management system. The geodatabase schema and the open application program interface (API) support six DBMS platforms: Oracle, Microsoft SQL Server, SQL Server Express, IBM DB2, PostgreSQL, and Informix. The DBMS server platform includes executables that enable query processing while spatial data is compressed within the geodatabase (roughly 50 percent compression). The data remains compressed during network transfer and is uncompressed by the client application, where it must go for analysis and display.

All ESRI client applications include the ArcSDE executables as part of a direct connect API that communicates directly with the DBMS network client software. (An SDE executable is a middleware translator, enabling ArcObjects to query the supported databases and manage the geodatabase schema.) These ArcSDE middleware functions are embedded within the client applications. The DBMS network client transmits data to the DBMS server, where the query processing takes place.

ArcSDE executables can be installed on a remote server or on the DBMS server platform. If installed on a remote server, the ArcSDE application server would use a DBMS connection to a local DBMS network client to enable communication to the DBMS platform. When the ArcSDE application server is installed on the same platform as the DBMS, it will use the server client libraries and spawn an ArcSDE client connection that will communicate directly with the DBMS server client connections.

Terminal client access protocols

ESRI customers use Windows Terminal Server to deploy centrally managed ArcGIS Desktop sessions for display and control by remote terminal clients. Two terminal client protocols are available to support the remote client desktop environment. Microsoft provides a remote desktop protocol (RDP) that works with Windows client platforms, and Citrix provides an independent computing architecture (ICA) protocol through implementation of its software. Both protocols implement compressed data transmissions that perform well over limited bandwidth connections.

The Citrix XenApp server (formerly Presentation Server) provides advanced server administration, security, and a variety of supported terminal client platforms beyond what is available with the Microsoft RDP client. Most ESRI customers include Citrix XenApp to provide access to Desktop GIS applications.

Web communication protocols

Many people have grown accustomed to easy access over the Web with HTTP (hypertext transfer protocol), a standard Web transmission protocol. It allows for publishing Web applications for "thin" browser clients (the curious and casual) and Web services for heavier desktop applications. In this transaction-based environment, the browser or the desktop client controls the application service selection and display. A Web application service provider publishes a map for browser clients to see. Browsers can zoom in on it and query it, but because the map is published based on the Web application, the size of the display traffic can be optimized for fast delivery. The ArcGIS Desktop display environment, however, is controlled by the client application. Traffic for the ArcGIS Desktop is higher because of the larger image transfers. Image size is proportional to the physical screen display size; thus, larger image displays can result in higher display traffic.

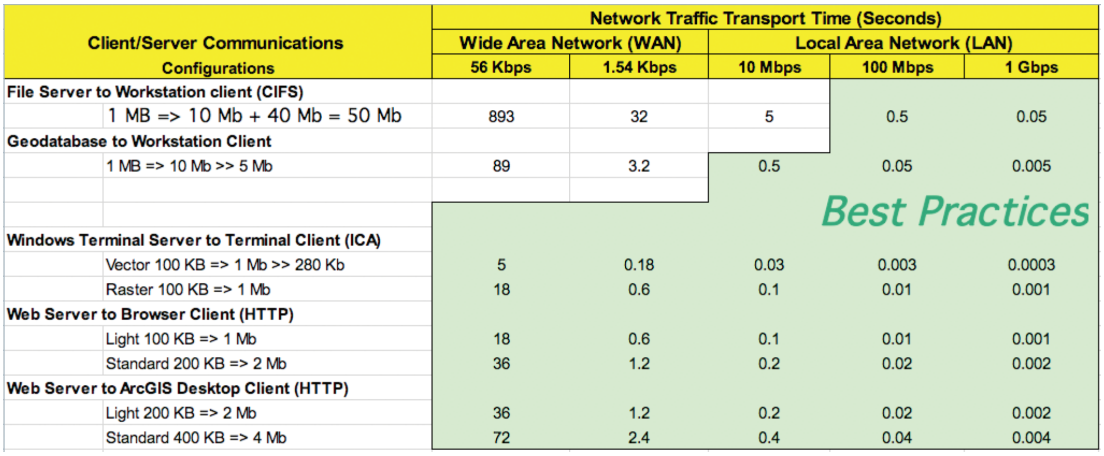
Network communication performance

Network communications can affect how the user experiences computer performance in several different ways. The primary and most obvious performance impact is data transport time: the time it takes to transfer the data for the client display. A typical GIS application requires up to 1 MB of data to generate each map display, and that translates to about 10 Mb of traffic (uncompressed) for the data alone. (For the entire network traffic per display, you must also account for the extra traffic required to process the data transmission.)

In figure 3-5, a simple traffic display analysis shows the minimum transport times over the network for the typical GIS communication protocols. Network traffic transport times are computed for a 56 Kbps dial-up connection, 1.54 Mbps T-1 WAN, and for 10 Mbps, 100 Mbps, and 1 Gbps local network connections. Tightly coupled client/server workflows are listed first on the left, followed by Windows Terminal Server and Web Services workflows. The chart demonstrates the importance of considering network infrastructure capacity in the overall system design: you want to be safe within the shaded area for best practices and reasonable transport times.

You can use the data transfer volume (traffic per display) and available network bandwidth to calculate the minimum network transport times for a single map display transaction, for various workflow configurations, as shown in the chart. For the client/server configuration at the top (file server to workstation client), each application display needs 1 MB of data, which is the equivalent of 10 Mb of traffic. Up to 40 Mb of additional traffic is generated for the file server access, bringing the total to 50 Mb of traffic per display. To come to a best-case estimate of the data transport time, you simply divide the total traffic required (50 Mb) by the network bandwidth available. In the case of our client/server configuration, you would need a 100 Mbps bandwidth to transfer display traffic in less than one second.

It is easy to see that client access to a file data source makes sense only in a LAN environment. (There,



Client/Server Communications Configurations	Network Traffic Transport Time (Seconds)				
	Wide Area Network (WAN)		Local Area Network (LAN)		
	56 Kbps	1.54 Kbps	10 Mbps	100 Mbps	1 Gbps
File Server to Workstation client (CIFS)					
1 MB => 10 Mb + 40 Mb = 50 Mb	893	32	5	0.5	0.05
Geodatabase to Workstation Client					
1 MB => 10 Mb >> 5 Mb	89	3.2	0.5	0.05	0.005
				Best Practices	
Windows Terminal Server to Terminal Client (ICA)					
Vector 100 KB => 1 Mb >> 280 Kb	5	0.18	0.03	0.003	0.0003
Raster 100 KB => 1 Mb	18	0.6	0.1	0.01	0.001
Web Server to Browser Client (HTTP)					
Light 100 KB => 1 Mb	18	0.6	0.1	0.01	0.001
Standard 200 KB => 2 Mb	36	1.2	0.2	0.02	0.002
Web Server to ArcGIS Desktop Client (HTTP)					
Light 200 KB => 2 Mb	36	1.2	0.2	0.02	0.002
Standard 400 KB => 4 Mb	72	2.4	0.4	0.04	0.004

Figure 3-5

GIS network transport time for display traffic.

workstation connections are 100 Mbps these days to optimize user productivity in a file-based environment.) Geodatabase access is about 10 times better than file access. Access over a dial-up connection still does not make sense; transport times are far too slow for most user workflows. Some customers use one or two geodatabase connections over a T-1 connection, although this is not recommended.

Terminal and browser clients access only the display environment over the network. Windows Terminal Server displays require about 100 KB of data, or about 1 Mb of traffic. Vector traffic is compressed to less than 280 Kb per display—image traffic does not compress as well. Windows Terminal Server clients provide the best user performance over limited bandwidth environments.

By using published Web applications to provide displays, you generate less traffic, thereby optimizing network performance. To do this, you need to be sensitive to the traffic required to provide each display. Consider that standard ArcIMS client displays tend to be around 100 KB in size. Most of the newer ArcGIS Server published applications and services are more like 200 KB per display. ArcGIS Desktop data services can be much higher, since they usually request an image service in order to produce a full, high-resolution ArcMap desktop display.

If possible, the most frequently accessed displays should be very simple, without lots of pictures or maps that require much network traffic. Opting for small and simple map displays reduces traffic transfer time, but using cached maps is even better. If real-time data is not involved and you can get by with displays or maps that are prerendered, caching is one of the best ways to optimize performance. As an example of displays, we've been talking mostly about dynamic maps here, ones produced on demand or "on the fly." But what if you didn't have to transfer all that data back and forth every time you wanted a map? For desktop clients, data makes one trip to the client and then is available in a local cache for follow-on displays. Cached maps and cached data make sense if the maps and data don't change too often; even caching a basemap containing only the static data will save network traffic and transport time.

As opposed to the typical remote client displays requiring 100 KB of data, the richer ArcGIS Server environments require more like 200 KB of data to provide a dynamic display, but the 9.2 and 9.3 versions provide a variety of data caching options to compensate for that and optimize performance. Static data can be preprocessed and provided as a client basemap file-based image service. The image files are delivered to a client cache during the initial display, so that follow-on displays can use the data now localized in a cache. Implementation of client data cache architecture can reduce the dynamic layer display requirements and network display traffic, using less server and network resources and improving client performance.

It is important to consider network traffic in your design analysis. During peak work periods, workflows can slow to a crawl similar to what is experienced in big city driving on major highway arteries during rush hour. Many remote client performance problems stem from just such a traffic jam scenario on the network. Sufficient bandwidth is critical to enabling user productivity during peak business hours.

Display traffic transfer times are only part of the overall network performance challenge. Other traffic on the network will also reduce available bandwidth. ArcGIS Desktop DBMS session-based workflows are tightly coupled, and unstable network connections can cause the client application to get out of sync with the DBMS server session. When this happens, the database should recover to the last consistent state (last save). This can be a very frustrating experience for database maintenance workflow operations or even GIS analysis or project work. Network latency is also a consideration, since geodatabase queries can require hundreds of sequential server communications to support a single map display.

You have seen the best practices identified for each of the available configuration alternatives. Distributed client/server applications perform best in a LAN environment. Remote desktop users should be supported from a Windows Terminal Server or distributed desktop environment (application runs in the computer room with local data access, and persistent remote client access is provided for active session display and control). Web services work fine over widely distributed WAN and Internet environments. Be sensitive to the published display traffic and clients will be more productive.

It may be necessary to upgrade the existing bandwidth. I was in Atlanta this past year to teach a class, and was staying at a hotel only five miles from the classroom. It took over an hour to drive to class during the morning rush hour—it was a parking lot the whole way. People like housing developments with pretty parks, swimming pools, and schools for the kids. Cities like to build technology centers for company offices, hotels, and restaurants away from family and children. Atlanta has both—but the roads connecting the housing community with the business park were never upgraded to support the traffic. Transportation bandwidth is an important design consideration.

Network latency

Several key performance factors contribute to overall display response time (chapter 7). These factors include display, network, and database processing time plus any queue time (processing delays) and network latency

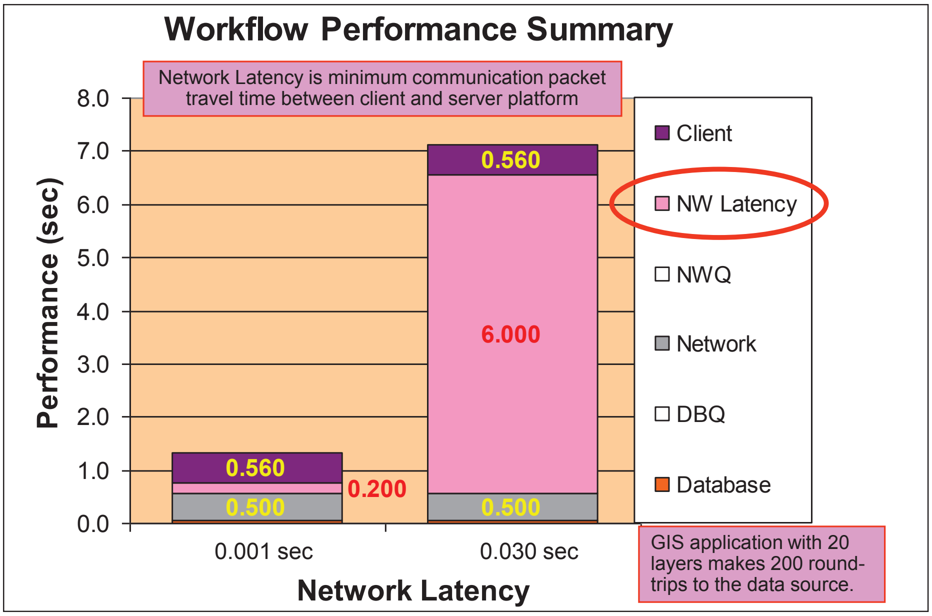
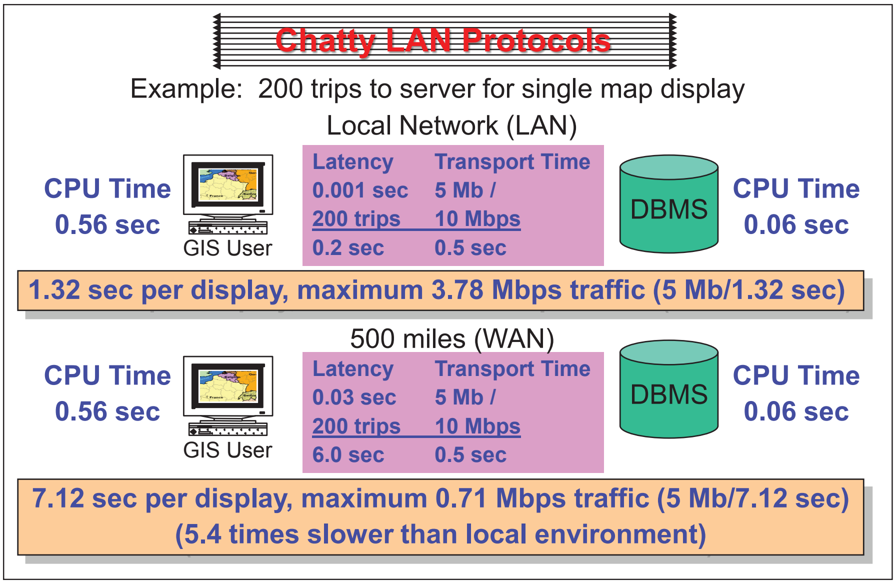


Figure 3-6

Network latency.



Chatty LAN Protocols

Example: 200 trips to server for single map display

Local Network (LAN)

CPU Time 0.56 sec

GIS User

Latency	Transport Time
0.001 sec	5 Mb /
200 trips	10 Mbps
0.2 sec	0.5 sec

DBMS

CPU Time 0.06 sec

1.32 sec per display, maximum 3.78 Mbps traffic (5 Mb/1.32 sec)

500 miles (WAN)

CPU Time 0.56 sec

GIS User

Latency	Transport Time
0.03 sec	5 Mb /
200 trips	10 Mbps
6.0 sec	0.5 sec

DBMS

CPU Time 0.06 sec

7.12 sec per display, maximum 0.71 Mbps traffic (5 Mb/7.12 sec)
(5.4 times slower than local environment)

Figure 3-7

Network latency considerations.

(medium travel time). Queue time is the duration when program instructions are waiting in line to be processed; these processing delays are caused by random arrival times. (We begin to experience processing delays when traffic reaches about 50 percent of network capacity, and the delay time continues to increase as network traffic increases beyond 50 percent capacity.) Network latency is the time it takes for a network packet to travel from client to server (travel time), while network transport time represents the processing time incurred at the network connection—time required to get the data on the network medium (based on bandwidth capacity). Network latency can be measured with a simple ping utility ("tracert" <trace route> DOS command is one example of a latency measurement), which identifies the hardware components (routers) in the transmission path and the travel time between each network connection.

Client/server protocols that make several trips back and forth between the client and server to complete processing for one display are called "chatty" transactions. Each round-trip will incur the network latency (packet travel time) delays. Figures 3-6 and 3-7 provide an example of network latency for a low latency and high latency connection, where bandwidth capacity over the network connection are both 10 Mbps. The complete client map display processing time is approximately 0.56 seconds. Network capacity is less than 50 percent, so there is no queue time in this example.

Database access protocols are “chatty”: a typical database query requires a large number of trips to and from the server to complete the client display transaction. The total number of trips to and from the server (query transactions) will vary based on the data model complexity (primarily based on the number of feature classes or layers in the display). Figure 3-7 demonstrates how network latency can make a difference even when there is plenty of network bandwidth.

What affects display response time the most? For LAN environments, network latency is very low (typically less than 0.001 milliseconds per trip to the server). Even numerous trips between the server and client don’t limit performance very much. Primarily, what determines how long you wait for the display are the client and server processing and network data transport times. In the example, computer processing time is 0.62 (0.56 + 0.06) seconds; network transport time is 0.5 (5 Mb/10 Mbps) seconds; and network latency is 0.2 (0.001 x 200) seconds. Total display response time is 1.32 seconds. Average network traffic is 3.78 (5 Mb/1.32) seconds. This traffic is well below 50 percent of the 10 Mbps network bandwidth capacity, so we would not expect queue time delays.

For WAN distances, which are longer and normally include multiple router hops, there can be measurable network latency delay. Network latency can have a considerable performance impact when using chatty database protocols. In the figure 3-7 example, the total transaction time over the WAN (including cumulative network latency) is 7.12 seconds. The maximum bandwidth used by a single user on this WAN connection is 0.71 Mbps, so you can see that user performance has been limited by network latency, not by WAN bandwidth. Many global WAN connections these days include satellite communication links. The fastest packet travel time is limited by the speed of light, which for very long distances (satellite connections) can result in network latency that is unacceptable. Good performance over WAN environments can be achieved from protocols that minimize sequential round-trips to the server (communication chatter). Latency is an important consideration when selecting remote client software solutions—good remote client solutions make use of Citrix Windows Terminal Server and Web software technology.

Shared network capacity

The total number of clients that can be supported on a single network segment (network backbone, server network interface card—NIC, campus network between buildings, etc.) is a function of network traffic transport time (amount of data traffic divided by network bandwidth) and the total number of concurrent clients. Only one client packet can be transmitted over a shared network segment at any time.

With older switch technology, multiple transmissions on the same Ethernet network segment would result in collisions. Recovery from a collision would require each client to send a repeat transmission to complete their packet delivery. Ethernet segments became quickly saturated when multiple collisions occurred because of the rapidly increasing number of transmissions. Ethernet switches today include a cache buffer (concurrent transactions wait in a cache until they can be transmitted over the network segment), which when configured properly can avoid network collisions and improve transmission efficiency.

Figure 3-8 shows multiple client sessions sharing the same network segment; each data exchange is represented by the small boxes. Only one data exchange can be supported at one time on the same network segment.

A GIS application can require 1 MB of spatial data, or up to 10 Mb of network traffic, to enable each map display. A 1 MB map display is illustrated in figure 3-9 (1:2,400 scale [feet], average features = 250). Applications can be tuned to prevent display of specific layers when the map extent exceeds defined thresholds. Reducing the number of layers underpinning the display improves performance. Only required data layers

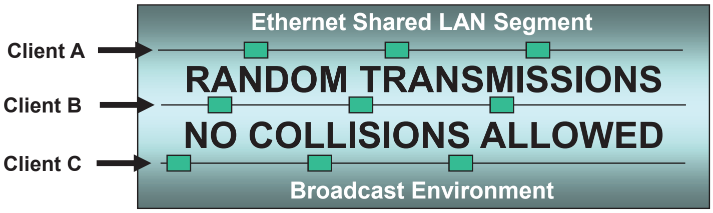


Figure 3-8

Shared network capacity.

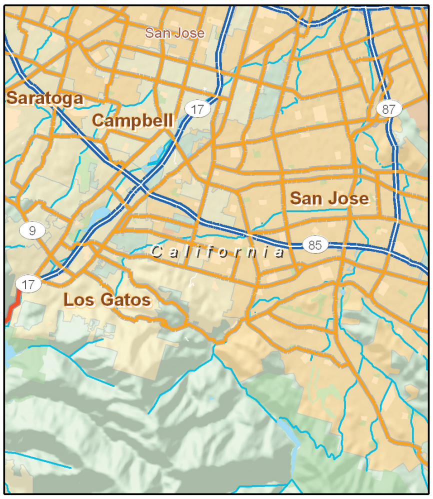


Figure 3-9

Typical 1 MB map display.

should be displayed for each map extent (e.g., it would not be appropriate to display individual parcel layers in a map showing the full extent of Kirkland, Washington). Proper tuning of the application can reduce network traffic and speed up performance.

Network configuration guidelines

Standard published guidelines are used for configuring network communication environments. These standards are GIS application–specific and based on typical user environment needs. Communication environments are statistical in nature, since only a percentage of user processing time requires transmission over the network. Network design standards represent typical GIS user workloads and address statistical application use, thereby providing the basis for the initial system design—general guidelines for establishing network requirements, as shown in figure 3-10.

The Capacity Planning Tool (chapter 10) will provide a flow analysis based on specific workflow traffic loads that address communication bandwidth requirements. Network data transfer time is a small fraction of the total display response time (on properly configured networks). Network data transfer can be the largest factor contributing to display response time when bandwidth is too small or when too many clients are on the same shared network segment.

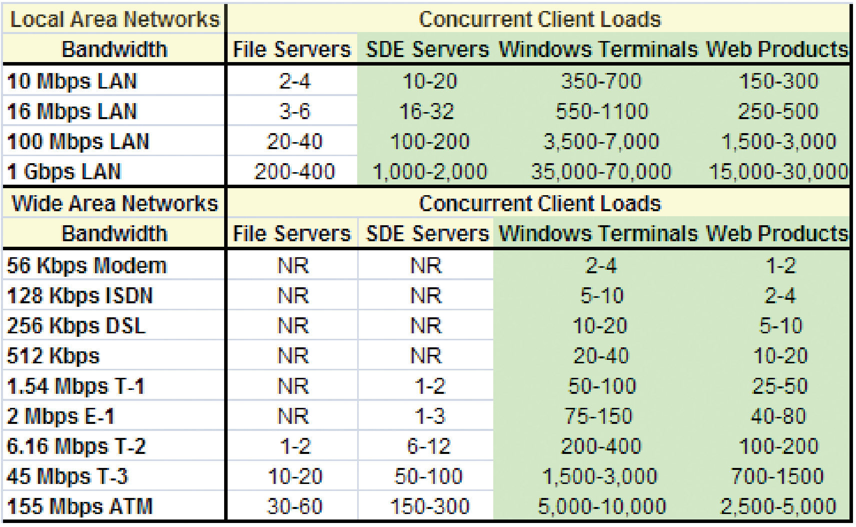
The network must be designed to support peak traffic demands. The amount of traffic varies based on the different types of applications and user work patterns. Standard guidelines provide a place to start configuring a network environment. Once the network is operational, network management becomes an ongoing traffic management task, strongly affected by the work environment and changes in computer technology. Network traffic demands should be monitored and necessary adjustments made to support peak user requirements.

Network design standards

Figure 3-10 provides recommended design guidelines to show you how big a bandwidth you will need to handle the network traffic generated by all the workflows you expect might be operating at once. These guidelines for network bandwidths establish preliminary guidelines for configuring distributed LAN and WAN workflows. Four separate GIS communication environments are included for each network bandwidth. The number of recommended clients is based on experience with actual system implementations and may not represent your worst-case scenarios. Networks should be configured with flexibility to provide special support to power users whose data transfer needs exceed typical GIS user bandwidth requirements.

Web services configuration guidelines

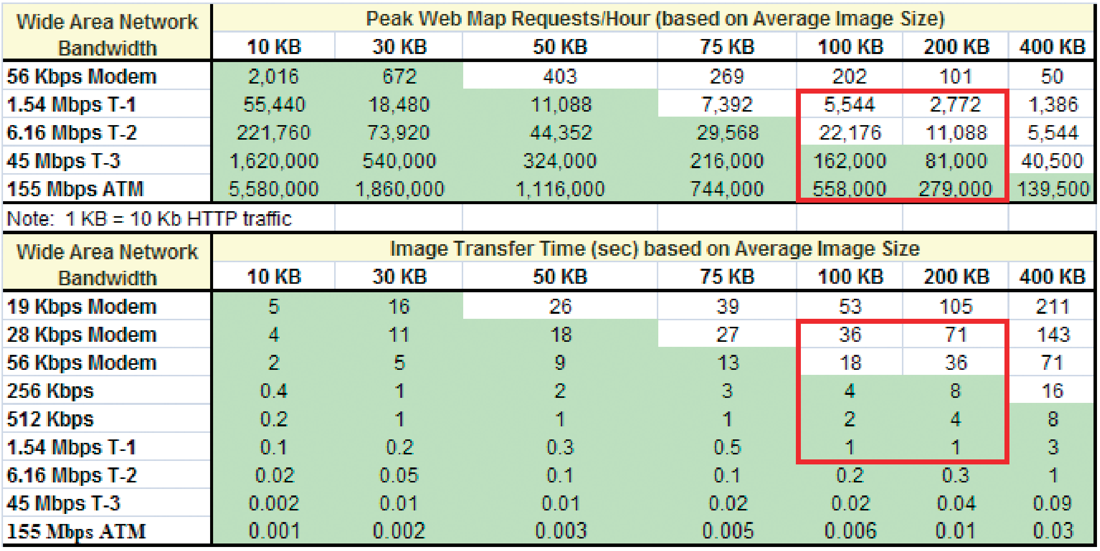
Implementation of Web mapping services places additional demands on the network infrastructure. The amount of system impact is related to the complexity of the published mapping services: Map services with small (less than 10 KB) or a limited number of complex images will have little effect on network traffic. Large



Local Area Networks	Concurrent Client Loads			
Bandwidth	File Servers	SDE Servers	Windows Terminals	Web Products
10 Mbps LAN	2-4	10-20	350-700	150-300
16 Mbps LAN	3-6	16-32	550-1100	250-500
100 Mbps LAN	20-40	100-200	3,500-7,000	1,500-3,000
1 Gbps LAN	200-400	1,000-2,000	35,000-70,000	15,000-30,000
Wide Area Networks	Concurrent Client Loads			
Bandwidth	File Servers	SDE Servers	Windows Terminals	Web Products
56 Kbps Modem	NR	NR	2-4	1-2
128 Kbps ISDN	NR	NR	5-10	2-4
256 Kbps DSL	NR	NR	10-20	5-10
512 Kbps	NR	NR	20-40	10-20
1.54 Mbps T-1	NR	1-2	50-100	25-50
2 Mbps E-1	NR	1-3	75-150	40-80
6.16 Mbps T-2	1-2	6-12	200-400	100-200
45 Mbps T-3	10-20	50-100	1,500-3,000	700-1500
155 Mbps ATM	30-60	150-300	5,000-10,000	2,500-5,000

Figure 3-10

Network design guidelines: Staying within the green is recommended.



Wide Area Network Bandwidth	Peak Web Map Requests/Hour (based on Average Image Size)						
	10 KB	30 KB	50 KB	75 KB	100 KB	200 KB	400 KB
56 Kbps Modem	2,016	672	403	269	202	101	50
1.54 Mbps T-1	55,440	18,480	11,088	7,392	5,544	2,772	1,386
6.16 Mbps T-2	221,760	73,920	44,352	29,568	22,176	11,088	5,544
45 Mbps T-3	1,620,000	540,000	324,000	216,000	162,000	81,000	40,500
155 Mbps ATM	5,580,000	1,860,000	1,116,000	744,000	558,000	279,000	139,500

Note: 1 KB = 10 Kb HTTP traffic

Wide Area Network Bandwidth	Image Transfer Time (sec) based on Average Image Size						
	10 KB	30 KB	50 KB	75 KB	100 KB	200 KB	400 KB
19 Kbps Modem	5	16	26	39	53	105	211
28 Kbps Modem	4	11	18	27	36	71	143
56 Kbps Modem	2	5	9	13	18	36	71
256 Kbps	0.4	1	2	3	4	8	16
512 Kbps	0.2	1	1	1	2	4	8
1.54 Mbps T-1	0.1	0.2	0.3	0.5	1	1	3
6.16 Mbps T-2	0.02	0.05	0.1	0.1	0.2	0.3	1
45 Mbps T-3	0.002	0.01	0.01	0.02	0.02	0.04	0.09
155 Mbps ATM	0.001	0.002	0.003	0.005	0.006	0.01	0.03

Figure 3-11

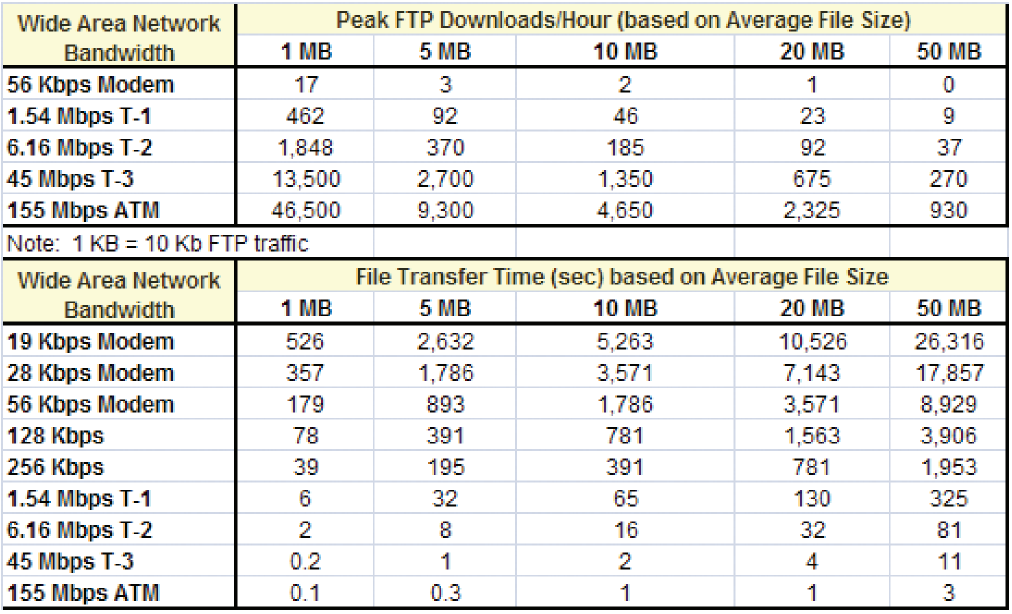
Web services network performance.

images (greater than 100 KB) can significantly impact network performance.

In figure 3-11, you'll see an overview of network performance characteristics you should consider when deploying a Web mapping solution. The top portion of the chart shows the maximum number of requests per hour that can be handled over various WAN bandwidths, based on average map image size. The bottom portion of the chart shows the optimum transmission time for various map images. You must design Web information products according to user performance needs. What's the available network bandwidth? That may be your primary performance consideration. Simple, high-performance map services can make do just fine with map images from 50 KB to 100 KB in size. Less traffic per display will minimize network transport time. (Consider that a 100 KB image uses up more than 36 seconds of network transport time for 28 Kbps clients, a response time most users would consider unreasonably slow.) Many Web developers overlook the fact that peak site capacity with an average of 100 KB display traffic means a maximum of 5,544 requests per hour over a single T-1 Internet service provider connection. You should ask the question, do we have enough bandwidth to support our peak transaction loads? A single entry-level ArcGIS Server platform can support up to 25,000 map requests per hour, requiring more bandwidth than most customer sites have for Internet access services. More complex ArcGIS Server map services generate 100 KB to 200 KB traffic per display. ArcGIS Desktop users may request image services from 200 KB to 400 KB in size (image size varies with user display size and resolution). Users generally demand reasonable performance or they will not be pleased with the service. Adequate network bandwidth capacity and careful information-product design are primary considerations when developing popular high-performance Web applications.

ArcGIS Server can make high-performance, complex services possible by using a preprocessed data cache. The more intelligent clients (ArcGIS Desktop, ArcGIS Explorer, Web applications with Adobe Flash clients, etc.) are able to overlay Web-based vector and image services on top of a high-performance, local, cache layer. The local cache data can be sent from the server once and used by the client many times, since the images are stored on the local client machine. Client data caching can reduce network transport times and increase display performance, when configured and used properly.

You can get data delivery as a service from the Web, downloading data to clients over the Internet. The data delivery service extracts data layers from the geodatabase based on identified extent, then zips the data into a compressed file, and downloads the data to the client. Standard Web-server, file-transfer applications can do similar things. Figure 3-12 identifies minimum download times based on available bandwidth and the size of compressed data transfers. Data downloads should be restricted to protect Web service bandwidth, as they can very easily consume all available bandwidth and slow down performance for other Web mapping clients. You need to control or limit the amount of traffic capacity that anyone can use for downloading data.

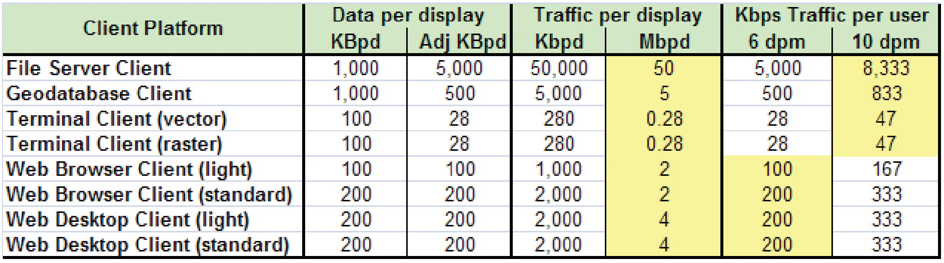


Wide Area Network Bandwidth	Peak FTP Downloads/Hour (based on Average File Size)				
	1 MB	5 MB	10 MB	20 MB	50 MB
56 Kbps Modem	17	3	2	1	0
1.54 Mbps T-1	462	92	46	23	9
6.16 Mbps T-2	1,848	370	185	92	37
45 Mbps T-3	13,500	2,700	1,350	675	270
155 Mbps ATM	46,500	9,300	4,650	2,325	930
Note: 1 KB = 10 Kb FTP traffic					

Wide Area Network Bandwidth	File Transfer Time (sec) based on Average File Size				
	1 MB	5 MB	10 MB	20 MB	50 MB
19 Kbps Modem	526	2,632	5,263	10,526	26,316
28 Kbps Modem	357	1,786	3,571	7,143	17,857
56 Kbps Modem	179	893	1,786	3,571	8,929
128 Kbps	78	391	781	1,563	3,906
256 Kbps	39	195	391	781	1,953
1.54 Mbps T-1	6	32	65	130	325
6.16 Mbps T-2	2	8	16	32	81
45 Mbps T-3	0.2	1	2	4	11
155 Mbps ATM	0.1	0.3	1	1	3

Figure 3-12

Data download performance.



Client Platform	Data per display		Traffic per display		Kbps Traffic per user	
	KBpd	Adj KBpd	Kbpd	Mbpd	6 dpm	10 dpm
File Server Client	1,000	5,000	50,000	50	5,000	8,333
Geodatabase Client	1,000	500	5,000	5	500	833
Terminal Client (vector)	100	28	280	0.28	28	47
Terminal Client (raster)	100	28	280	0.28	28	47
Web Browser Client (light)	100	100	1,000	2	100	167
Web Browser Client (standard)	200	200	2,000	2	200	333
Web Desktop Client (light)	200	200	2,000	4	200	333
Web Desktop Client (standard)	200	200	2,000	4	200	333

Figure 3-13

Network design planning factors. The yellow-highlighted traffic per display (Mbpd) figures above are used as network load factors in the Capacity Planning Tool in chapter 10.

Network planning factors

Many network administrators establish and maintain metrics on network use. These metrics help them estimate increased network demands as they plan for future user deployments. Figure 3-13 identifies standard network design planning factors for typical GIS clients (highlighted in yellow), based on their target data source. You will be seeing these numbers again in later chapters as you learn to project network bandwidths adequate to GIS user needs in planned GIS deployments.

Network traffic delays can have a significant effect on user response times. Ever wonder what happens when you click and then . . . nothing? It might be all those intricate and beautiful maps—higher display traffic requirements—that are consuming all available network bandwidth resources. Delays begin to occur once traffic exceeds 50 percent of the bandwidth capacity, determined by the weakest network connection. Network latency can be a major concern with chatty client/server protocols (desktop applications accessing database or file data sources).

Network protocols enable distributed software communications. The next chapter will take a look at all the different ways GIS software can be distributed to support your enterprise operations. Now that you know what these communication protocols are, and what they mean in regard to network traffic and performance, you are prepared to review these communication options and select the best platform architecture to support your GIS needs.