CHAPTER 7

TCP/IP Applications

7.1 INTRODUCTION

TCP/IP applications are the applications which use TCP/IP protocols in order to function. The Internet model uses Client-Server architecture. There are various applications based on the internet Model and these generally use the application layer of the TCP/IP reference model. These applications and services are generally of two types:

- The applications which a user uses directly such as the Telnet and Email
- The applications which are not used directly by a user. These applications provide support to other applications and services. Examples include the Domain Name System.

These applications are summarized as follows :

- **Bootstrap Protocol**: Bootstrap Protocol (BOOTP) is an application used for providing a dynamic method for connecting workstations with servers. It also makes it provides a dynamic method for assigning workstation Internet Protocol (IP) addresses and initial program load (IPL) sources.
- **Internet Setup Wizard:** connecting the system to the internet (ISP) is possible via the use of Internet Setup Wizard. Eg. It can be used to set up a computer system as a web data server or application server.
- **Dynamic Host Configuration Protocol:** Dynamic Host Configuration Protocol (DHCP) is a TCP/IP Protocol which uses a central server to manage all the IP addresses and other configuration details of an entire network.
- **Domain Name System:** Domain Name System (DNS) is a hierarchical distributed database system which is responsible for managing host names and their associated Internet Protocol (IP) addresses.
- **E-mail:** This application is used to configure, use, manage, and troubleshoot e-mails on the system.
- **File Transfer Protocol:** FTP is an acronym for File Transfer Protocol. As the name suggests, FTP is used to transfer files between computers on a network. FTP may be used to exchange files between computer accounts, transfer files between an account and a desktop computer, or access online software archives.
- **IP filtering and network address translation:** IP filtering and network address translation (NAT) work like a firewall to protect the network from intruders.
- **HTTP Server** HFS (HTTP File Server) is an application which is used to send and receive files. It's different from classic file sharing because it uses web technology to be more compatible with today's

Internet. It also differs from classic web servers because it's very easy to use and runs "right out-of-the box".

- **Remote Access Services: PPP connections**: Point-to-Point Protocol (PPP) is an Internet standard for transmitting data over serial lines.
- **Remote Execution:** The Remote Execution (REXEC) server is a Transmission Control Protocol/ Internet Protocol (TCP/IP) application that allows a client user to submit system commands to a remote system. The Remote Execution Protocol (REXEC) allows processing of these commands or programs on any host in the network. The local host then receives the results of the command processing.
- **Telnet**: Telnet is a protocol that enables the user to log onto and the user at the terminal interacts with the local telnet client. The TELNET client acts as a normal terminal accepting any keystrokes from the keyboard, interpreting them and displaying the output on the screen. The client on the computer makes the TCP connection to the host machine's port 23 where the TELNET server answers. The TELNET server interacts with applications in the host machine and assists in the terminal emulation.
- **Trivial File Transfer Protocol:** Trivial File Transfer Protocol (TFTP) is a simple protocol that provides basic file transfer function with no user authentication.
- **Virtual private networking:** A virtual private network (VPN) allows a company to securely extend its private intranet over the existing framework of a public network, such as the Internet. With VPN, a company can control network traffic while providing important security features such as authentication and data privacy.

7.2 VOIP

VoIP or Voice over Internet Protocol allows a user to receive telephone calls over the internet. It is also called as the "IP Telephony". A requirement for a VoIP connection is a high speed internet connection, such as cable, DSL, T1, or even wireless, as long as the wireless provider's bandwidth meets the minimum requirements for the VoIP provider you choose and their connection is stable enough to offer a connection without packet loss and jitter.

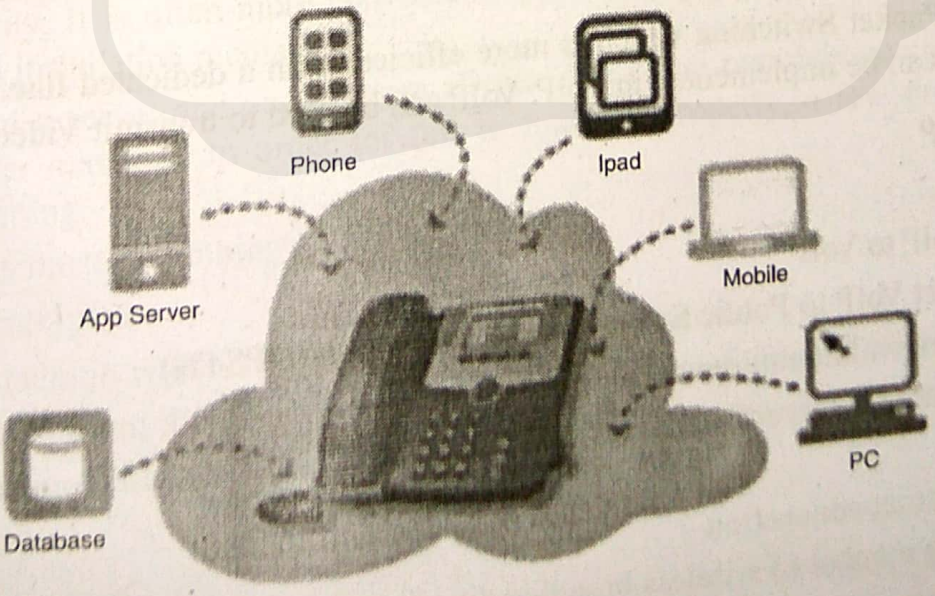

Fig. 7.1. VoIP

- VoIP allows you to make telephone calls using a computer network, over a data network like the Internet. VoIP converts the voice signal from your telephone into a digital signal that travels over the internet then converts it back at the other end so you can speak to anyone with a regular phone number. When placing a VoIP call using a phone with an adapter, you'll hear a dial tone and dial just as you always have. VoIP may also allow you to make a call directly from a computer using a conventional telephone or a microphone.
- Voice conversations are turned into digitized data and packetized for transmission across a network.

7.2.1 Working

- VoIP looks for IP address and translates phone numbers to IP addresses
- The **central call processor** is a piece of hardware running a specialized database/mapping program called a **soft switch**.
- Soft switches know:
  - Where the endpoint is on the network
  - What phone number is associated with that endpoint
  - The current IP address assigned to that endpoint
- If soft switch does not have the information, the request is handled by another soft switch.
- Used to connect different pieces of hardware.

7.2.2. Protocols Used

- H.323: This is the most widely used protocol which provides specifications for real-time, interactive videoconferencing, data sharing and audio applications (VoIP)
- SIP: Session Initiation Protocol is a more streamlined protocol which was developed specifically for VoIP. It is used for signaling and controlling multimedia communication sessions.

VoIP VS PSTN

- PSTN = Public Switched Telephone Network require a dedicated line to connect to the network. Compression algorithms cannot be implemented in PSTN. Also it is not possible to send video data over traditional PSTN.
- VoIP uses Packet Switching which is more efficient than a dedicated line. Also compression algorithms can be implemented in VoIP. VoIP can be used to transmit Video data as well.

Advantages of VoIP

- Cost
  - Free VoIP to VoIP
  - Low cost VoIP to Public Switch Telephone Network (PSTN)
  - Less bandwidth requirements
  - Low cost / no cost software and hardware
- Mobility
  - Any internet connection
  - Growing number of wireless broadband locations

Disadvantages of VoIP

- Quality
  - High quality PSTN
  - Variable VoIP dependent on connection
- Dependent on wall power
- Lost or delayed packets cause drop-out in voice
- Emergency Calls
  - Hard to find geographic location
- Security
  - Most VoIP services do not support encryption

7.3 NFS

In 1984 Sun Microsystems came up with Network File System (NFS), which allows a server to share directories and files with clients over a network. With NFS, users and programs can access files on remote systems as if they were stored locally.

NFS has many practical uses. Some of the more common uses include:

- Data that would otherwise be duplicated on each client can be kept in a single location and accessed by clients on the network.
- Several clients may need access to the /usr/ports/distfiles directory. Sharing that directory allows for quick access to the source files without having to download them to each client.
- On large networks, it is often more convenient to configure a central NFS server on which all user home directories are stored. Users can log into a client anywhere on the network and have access to their home directories.
- Administration of NFS exports is simplified. For example, there is only one file system where security or backup policies must be set.
- Removable media storage devices can be used by other machines on the network. This reduces the number of devices throughout the network and provides a centralized location to manage their security. It is often more convenient to install software on multiple machines from a centralized installation media.

NFS consists of a server and one or more clients. The client remotely accesses the data that is stored on the server machine. In order for this to function properly, a few processes have to be configured and running.

These daemons must be running on the server:

Daemon	Description
nfsd	The NFS daemon which services requests from NFS clients.
mountd	The NFS mount daemon which carries out requests received from nfsd.
rpcbind	This daemon allows NFS clients to discover which port the NFS server is using.

7.4 TELNET

Telnet is a protocol that allows you to connect to remote computers (called hosts) over a TCP/IP network (such as the Internet). Using telnet client software on your computer, you can make a connection to a

telnet server (i.e., the remote host). Once your telnet client establishes a connection to the remote host, your client becomes a virtual terminal, allowing you to communicate with the remote host from your computer. In most cases, you'll need to log into the remote host, which requires that you have an account on that system. Occasionally, you can log in asguest or public without having an account.

Telnet clients are available for all major operating systems.

Command-line telnet clients are built into most versions of Mac OS X, Windows, Unix, and Linux. To use these clients, go to their respective command lines (i.e., the Terminal application in Mac OS X, the shell in Unix or Linux, or the DOS prompt in Windows), and then enter:

Telnet host port

Replace host with the address of the service, and port with the port number on which the service runs (e.g., 80for http).

- TELNET is a general protocol, meant to support logging in from almost any type of terminal to almost any type of computer.
- It allows a user at one site to establish a TCP connection to a login server or terminal server at another site.
- A TELNET server generally listens on TCP Port 23.

How it works

- A user is logged in to the local system, and invokes a TELNET program (the TELNET client) by typing
- telnet xxx.xxx.xxx

  where xxx.xxx.xxx is either a host name or an IP address.
- The TELNET client is started on the local machine (if it isn't already running). That client establishes a TCP connection with the TELNET server on the destination system.
- Once the connection has been established, the client program accepts keystrokes from the user and relays them, generally **one character at a time**, to the TELNET server.
- The server on the destination machine accepts the characters sent to it by the client, and passes them to a terminal server.
- A "terminal server" is just some facility provided by the operating system for entering keystrokes from a user's keyboard.
  - The terminal server treats the remote user as it would any other user logged in to the system, including relaying commands to other applications.
  - The terminal server passes outputs back to the TELNET server, which relays them to the client, which displays them on the user's screen.
- In general, a TELNET server is implemented as a master server with some number of slave servers. The master server listens for service requests from clients. When it hears one, it spawns a slave server to handle that specific request, while the master goes back to listening for more requests.
- The only thing that makes TELNET hard to implement is the heterogeneity of the terminals and operating systems that must be supported. Not all of them use the same control characters for the same purposes.

- To accomodate this heterogeneity, TELNET defines a Network Virtual Terminal (NVT). Any user TELNETting in to a remote site is deemed to be on an NVT, regardless of the actual terminal type being used.
- It is the responsibility of the client program to translate user keystrokes from the actual terminal type into NVT format, and of the server program to translate NVT characters into the format needed by the destination host. For data sent back from the destination host, the translation is the reverse.
- NVT format defines all characters to be 8 bits (one byte) long. At startup, 7 bit US ASCII is used for data; bytes with the high order bit = 1 are command sequences.
- The 128 7-bit long US ASCII characters are divided into 95 printable characters and 33 control codes. NVT maps the 95 printable characters into their defined values - decimal 65 = "A", decimal 97 = "a", etc.
- The 33 control codes are defined for NVT as:

ASCII Code	Decimal value	Meaning
NUL	0	NO - OP
BEL	7	Ring "terminal bell"
BS	8	Backspace; move cursor left
HT	9	Horizontal tab; move cursor right
LF	10	Line feed; move down one line; stay in same column
VT	11	Vertical tab; move cursor down
FF	12	`Form Feed
CR	13	Carriage return; move cursor to beginning of current line
all others		NO - OP

- NVT defines end-of-line to be a CR-LF combination - the two-character sequence.
- In addition to the 128 characters mentioned above, there are 128 other possible characters in an 8-bit encoding scheme. NVT uses these 128 (with decimal values 128 through 255, inclusive) to pass control functions from client to server.

TELNET Operation

- The TELNET protocol is based on three ideas:
  - The Network Virtual Terminal (NVT) concept. An NVT is an imaginary device having a basic structure common to a wide range of real terminals. Each host maps its own terminal characteristics to those of an NVT, and assumes that every other host will do the same.
  - A symmetric view of terminals and processes .
  - Negotiation of terminal options. The principle of negotiated options is used by the TELNET protocol, because many hosts wish to provide additional services, beyond those available with the NVT. Various options may be negotiated. Server and client use a set of conventions to establish the operational characteristics of their TELNET connection via the "DO, DON'T, WILL, WON'T" mechanism discussed later in this document.

- The two hosts begin by verifying their mutual understanding. Once this initial negotiation is complete, they are capable of working on the minimum level implemented by the NVT.
- After this minimum understanding is achieved, they can negotiate additional options to extend the capabilities of the NVT to reflect more accurately the capabilities of the real hardware in use.
- Because of the symmetric model used by TELNET, both the host and the client may propose additional options to be used.
- The set of options is not part of the TELNET protocol, so that new terminal features can be incorporated without changing the TELNET protocol (mouse?).
- All TELNET commands and data flow through the same TCP connection.
- Commands start with a special character called the Interpret as Command escape character (IAC).
- The IAC code is 255.
- If a 255 is sent as data - it must be followed by another 255
- Each receiver must look at each byte that arrives and look for IAC. If IAC is found and the next byte is IAC - a single byte is presented to the application/terminal.
- If IAC is followed by any other code - the TELNET layer interprets this as a command.

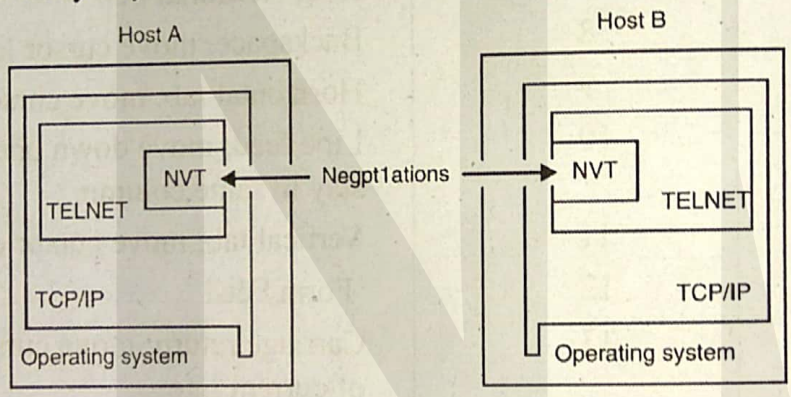

Fig. 7.2. Telnet Architecture

TELNET Options

- There is an extensive set of TELNET options, and the reader should consult tho Official Internet Protocol Standards for the standardization state and status for each of them.
- At the time of writing, the following options were defined:

Num	Name	State	RFC	STD
255	Extended-Options-List	Standard	861	32
0	Binary Transmission	Standard	856	27
1	Echo	Standard	857	28
3	Suppress Go Ahead	Standard	858	29
5	Status	Standard	859	30
6	Timing Mark	Standard	860	31
34	Linemode	Draft	1184	
2	Reconnection	Proposed		

(Contd...)

Num	Name	State	RFC	STD
4	Appr ox Message Size Negotiation	Proposed		
7	Remote Controlled Trans and Echo	Proposed	726	
8	Output Line Width	Proposed		
9	Output Page Size	Proposed		
10	Output Carriage-Return Disposition	Proposed	652	
11	Output Horizontal Tabstops	Proposed	653	
12	Output Horizontal Tab Disposition	Proposed	654	
13	Output Formfeed Disposition	Proposed	655	
14	Output Vertical Tabstops	Proposed	656	
15	Output Vertical Tab Disposition	Proposed	657	
16	Output Linefeed Disposition	Proposed	658	
17	Extended ASCII	Proposed	698	
18	Logout	Proposed	727	
19	Byte Macro	Proposed	735	
20	Data Entry Terminal	Proposed	1043	
21	SUPDUP	Proposed	736	
22	SUPDUP Output	Proposed	749	
23	Send Location	Proposed	779	
24	Terminal Type	Proposed	1091	
25	End of Record	Proposed	885	
26	TACACS User Identification	Proposed	927	
27	Output Marking	Proposed	933	
28	Terminal Location Number	Proposed	946	
29	TELNET 3270 Regime	Proposed	1041	
30	X.3 PAD	Proposed	1053	
31	Negotiate About Window Size	Proposed	1073	
32	Terminal Speed	Proposed	1079	
33	Remote Flow Control	Proposed	1372	
35	X Display Location	Proposed	1096	
39	TELENT Environment Option	Proposed	1572	
37	TELENT Authentication Option	Experimental	1416	

TELNET Command Structure

- The communication between client and server is handled with internal commands, *which are not accessible by users.*

- All internal TELNET commands consist of 2 or 3-byte sequences, depending on the command type.
- The Interpret As Command (IAC) character is followed by a command code. If this command deals with option negotiation, the command will have a third byte to show the code for the referenced option.

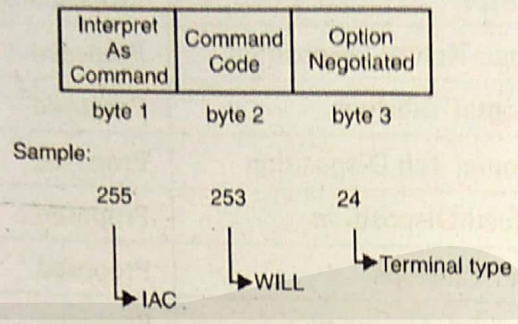

Fig. 7.3. Telnet Command Structure

- The key point is that, in order for any byte to be accepted as a command, it must be preceded by a byte with value 255 - "Interpret as Command" (IAC). Otherwise, the server program assumes that that byte is simply data that will be understood by the application program.
- This command proposes negotiation about terminal type.

7.5 FTP, SMTP, SNMP

FTP : File transfer protocol is used to transfer (upload/download) data from one computer to another over the internet or through or computer network. FTP is a most commonly communication protocol for transferring the files over the internet. Typically, there are two computers are involved in the transferring the files a server and a client. The client computer that is running FTP client software such as FileZila and SmartFTP etc initiates a connection with the remote computer (server). After successfully connected with the server, the client computer can perform a number of the operations like downloading the files, uploading, renaming and deleting the files, creating the new folders etc. Virtually operating system supports FTP protocols.

FTP Model (RFC 959)

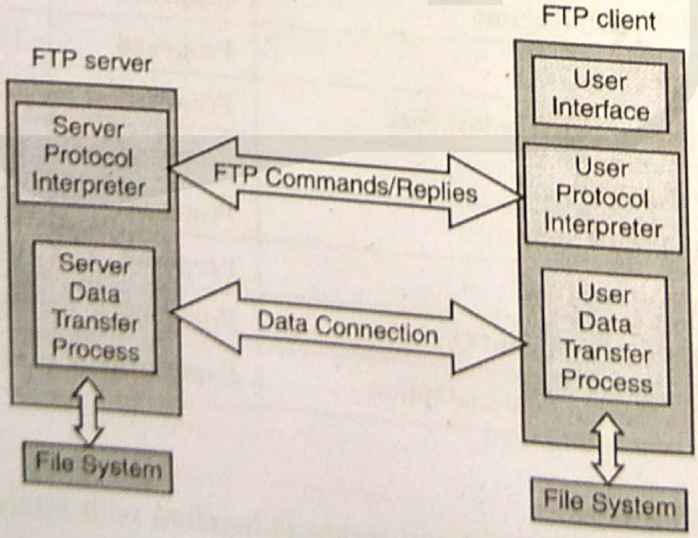

Fig. 7.4. FTP Model

7.5.1. SMTP: Simple Mail Transfer Protocol

SMTP is based on end-to-end delivery: An SMTP client contacts the destination host's SMTP server directly, on well-known port 25, to deliver the mail. It keeps the mail item being transmitted until it has been successfully copied to the recipient's SMTP. This is different from the store-and-forward principle that is common in many mailing systems, where the mail item can pass through a number of intermediate hosts in the same network on its way to the destination and where successful transmission from the sender only indicates that the mail item has reached the first intermediate hop. In various implementations, it is possible to exchange mail between the TCP/IP SMTP mailing system and the locally used mailing systems. These applications are called mail gateways or mail bridges. Sending mail through a mail gateway can alter the end-to-end delivery specification, because SMTP only guarantees delivery to the mail-gateway host, not to the real destination host located beyond the TCP/IP network. When a mail gateway is used, the SMTP end-to-end transmission is host-to-gateway, gateway-to-host, or gateway-to-gateway; the behavior beyond the gateway is not defined by SMTP. The agents used in SMTP may be grouped as:

- **MUA - Mail User Agent**
  - Users's email program
  - Used read and write email
- **MTA - Mail Transfer Agent**
  - Receives mail from MUA
  - Transfers mail to other MTA's
  - Transfers mail to receiver's MUA on request
- **SMTP - Simple Mail Transfer Protocol**
  - Use to send mail
    - From MUA to MTA
    - From MTA to MTA

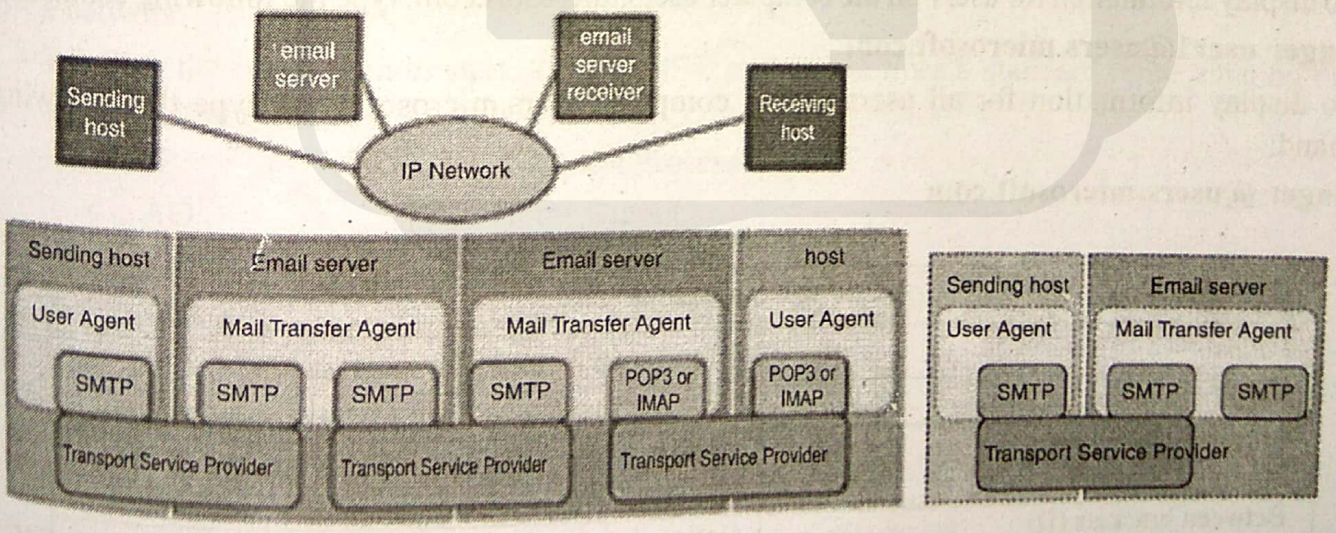

Fig. 7.5. SMTP Structure

SNMP: Simple Network Management Protoco is protocol used for the management of the network and is implemented by the network manager at a management station. Managed nodes consist of routers, hosts, etc. It must run SNMP agent which in turn keep objects (statistics etc).

Object defined in Management Information Base (MIB). Communication is with binary values Structure of binary values is defined in ASN.1 (Abstract Syntax Notation One) thus ASN.1 defines types of objects (int, string, arrays, etc.) and a standard binary representation of these.

7.6 FINGER

Displays information about a user or users on a specified remote computer (typically a computer running UNIX) that is running the Finger service or daemon. The remote computer specifies the format and output of the user information display. Used without parameters, finger displays help.

Syntax

finger [-l] [User] [@host] [...]

Parameters

-l : Displays user information in long list format.

User : Specifies the user about which you want information. If you omit the User parameter, finger displays information about all users on the specified computer.

@ host : Specifies the remote computer running the Finger service where you are looking for user information. You can specify a computer name or IP address.

/? : Displays help at the command prompt.

Remarks

- Multiple *User@host* parameters can be specified.
- You must prefix **finger** parameters with a hyphen (-) rather than a slash (/).
- This command is available only if the **Internet Protocol (TCP/IP)** protocol is installed as a component in the properties of a network adapter in Network Connections
- Windows 2000 and Windows XP do not provide a finger service.

Examples

To display information for user1 on the computer users.microsoft.com, type the following command:

finger user1@users.microsoft.com

To display information for all users on the computer users.microsoft.com, type the following command:

finger @users.microsoft.com

Formatting legend

Format	Meaning
Italic	Information that the user must supply
Bold	Elements that the user must type exactly as shown
Ellipsis (...)	Parameter that can be repeated several times in a command line
Between brackets ([])	Optional items
Between braces ({}); choices separated by pipe (|). Example: {even|odd}	Set of choices from which the user must choose only one
Courier font	Code or program output

Command name	Code	Comments
SE	240	End of sub-negotiation parameters.
NOP	241	No operation.
Data Mark	242	The Data stream portion of a synch. This should always be accompanied by a TCP urgent notification.
Break	243	NVT character BRK.
Go ahead	249	The GA signal.
SB	250	Indicates that what follows is sub-negotiation of the option indicated by the immediately following code.
WILL	251	Shows the desire to use, or confirmation that you are now using, the option indicated by the code immediately following.
WONT	252	Shows the refusal to use, or to continue to use, the option indicated by the code immediately following.
DO	253	Requests that the other party uses, or confirms that you are expecting the other party to use, the option indicated by the code immediately following.
DONT	254	Demands that the other party stop using, or confirms that you are no longer expecting the other party to sue, the option indicated by the code immediately following.
IAC	255	Interpret As Command. Indicates that what follows is a TELENT command, not data.

TELNET Basic Commands

- The primary goal of the TELNET protocol is the provision of a standard interface for hosts over a network.
- To allow the connection to start, the TELNET protocol defines a standard representation for some functions:
  - IP Interrupt Process
  - AO Abort Output
  - AYT Are You There
  - EC Erase Character
  - EL Erase Line
  - SYNCH Synchronize

  QUIT quit session

7.7 WWW, IPV6 AND NEXT GENERATION NETWORKS

WWW: World Wide Web and Its Component

The Web

An infrastructure of information combined and the network software used to access it

Web page
A document that contains or references various kinds of data

Links
A connection between one web page and another

Website
A collection of related web pages

Web browser
A software tool that retrieves and displays eb pages

Web server
A computer set up to respond to requests for web pages

Uniform Resource Locator (URL)
A standard way of specifying the location of a Web page, containing the hostname, "/", and a file

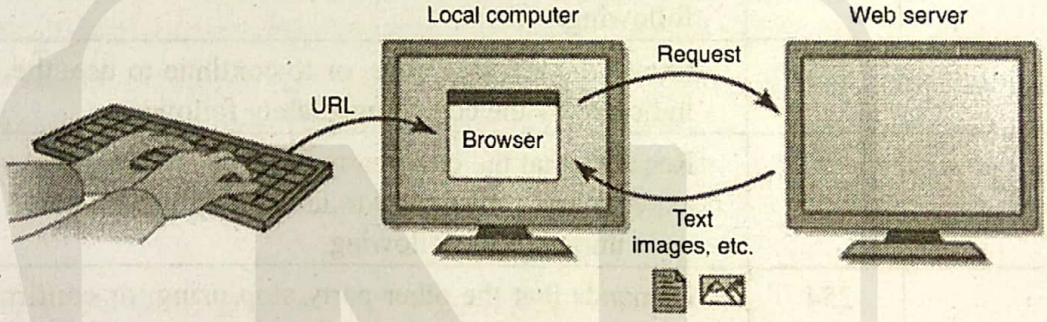

Fig. 7.6. URL Processing

7.7.1 Search Engine

A website that helps you find other websites

Instant messaging (IM)

- Applications that allow people to send short messages
- Similar to texting, but based on username not cellular phone number
- Some applications allow more than two users in a chat room
- If participants run application simultaneously, they can have an interactive conversation
- Most applications use proprietary protocols that dictate the precise format and structure of the messages
- Most instant messages are **not** secure

Cookie
A small text file that a web server stores on your local computer's hard disk

- A cookie contains information about your visit to the site
- Cookies can be used
  - to determine number of unique visitors to the site
  - to customize the site for future visits
  - to implement shopping carts that can be maintained from visit to visit

Cookies are **not** dangerous

Web analytics

- Collection and analysis of data regarding website usage
- Typically used by website owners to track the number and behavior of users visiting their sites
- **Example:** Google Analytics tracks
  - Where users are geographically located (based on their ISP)
  - Which site referred them
  - Which pages within your site they visit
  - How long they spend on each page
  - Which page they leave your site from

7.7.2 IPv6 – The next generation of IP

"In the general sense, an internet is a computer network that connects several networks. The Internet is a publicly available internationally interconnected system of computers plus the information and services provided to their users using a TCP/IP suite of packet switching communications protocols[1]".

To interconnect two or more computer networks it is necessary to have a routing device to exchange traffic, and steer traffic via several different nodes on the path across a network to its destination. The devices used to interconnect different networks are routers. Others devices with specific functions like gateways or bridge are also used. All network elements such as routers, switches, gateways, bridges, LAN cards, need to have at least one IP address.

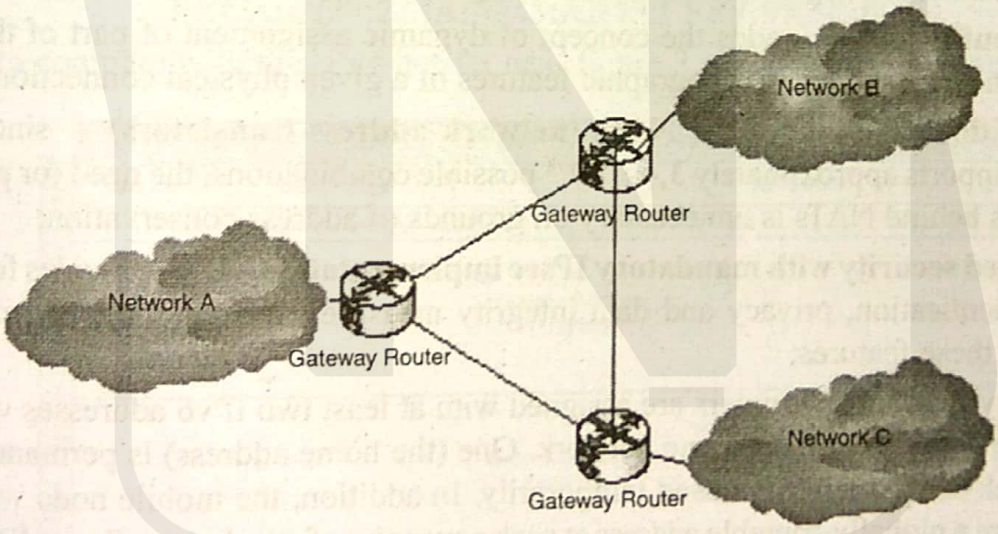

Fig. 7.7. Use of routers

Different IP packet networks are normally interconnected by Routers that have added functionality to permit accounting between the interconnected networks. In other configurations they act also as interworking devices between different protocols.

7.7.3 Major benefits of the IPv6 – Why change?

The new version of IPv6 was conceived to replace the previous IPv4 standard that was adopted two decades ago as a robust, easily implemented standard.

However IPv4 is being used successfully to support the communications systems in the emerging information society and has been updated to extend its useful life (e.g. NAT mechanism, IPsec protocol), MPLS, Tunnelling). However its capabilities are somewhat limited in the following areas:

- Exhaustion of the IPv4 address space;
- Growth of the Internet and the maintenance of routing tables
- Auto-configuration
- Mobility
- Security
- Quality of service

and the purpose of developing |IPv6 is to overcome these limitations.

The areas where IPv6 offers improvement are:

- **Expansion capacity for addressing and routing** – the IP address space is expanded from 32 bits to 128 bits, enabling a greatly increased number of address combinations, levels of hierarchical address organization and auto-configuration of addresses;
- **Simplified header format** – the IPv6 basic header is only 40 bytes long in spite of the greatly increased address allocation;
- **Enhanced options support** – several different, separate "extension headers" are defined, which enable flexible support for options without all of the header structure having to be interpreted and manipulated at every router point along the way;
- **Quality of service** – the Flow Label and the Priority fields in the IPv6 header are used by a host to identify packets that need special handling by IPv6 routers, such as non-default quality of service or "real-time" service. This capability is important in that it needs to support applications that require some degree of consistent throughput, delay, and jitter;
- **Auto-configuration** – adds the concept of dynamic assignment of part of the address space, based on geographic and topographic features of a given physical connection
- **Elimination of the need for NATs (network address translators)** – since the IP address space supports approximately 3.4×10^{38} possible combinations, the need for private addressing schemes behind NATs is unnecessary on grounds of address conservation;
- **Improved security with mandatory IPsec implementation** – IPv6 provides for integral support for authentication, privacy and data integrity measures, by requiring all implementations to support these features;
- **Mobility** - mobile computers are assigned with at least two IPv6 addresses whenever they are roaming away from their home network. One (the home address) is permanent; the other (the IPv6 link-local address) is used temporarily. In addition, the mobile node will typically auto-configure a globally-routable address at each new point of attachment. Every IPv6 router supports encapsulation, so every router is capable of serving as a home agent on the network(s) to which it is attached.

7.7.4 IP addressing architecture

An IP address is a binary number, which identifies any user's computer directly connected to the Internet. An IPv4 address consists of 32 bits, but it is usually represented by a group of four numbers (8 bits hexadecimal), from 0 to 255 ranges and separated by full stops. An example of this representation is showed bellow:

124.32.43.4

Several domain names can also be linked to the same IP address, in effect similar to having more than one name for the same person. The format of the IPv4 header is showed in figure 2:

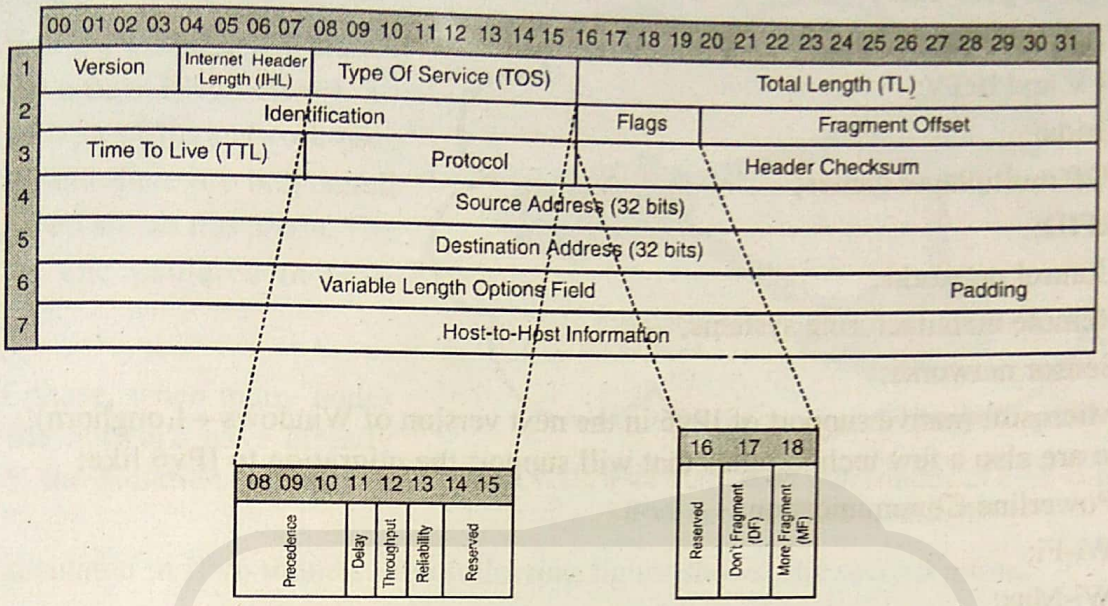

Fig. 7.8. IPv4 Structure

The most recognized change from IPv4 to IPv6 is the length of network addresses. The IPv6 addresses have 128 bits length. The 128 bits provide approximately $3.4x10^{38}$ separate values. An IPv6 address consists of eight numbers in the hexadecimal format, from 0 to 65535 (decimal) ranges and separated by a colon ":". An example of this new representation is showed following:

FECA:0000:234A:0043:AB45:FFFF:9A3E:000B

In other to compare with the IPv4 header next figure 3 shows the IPv6 format header:

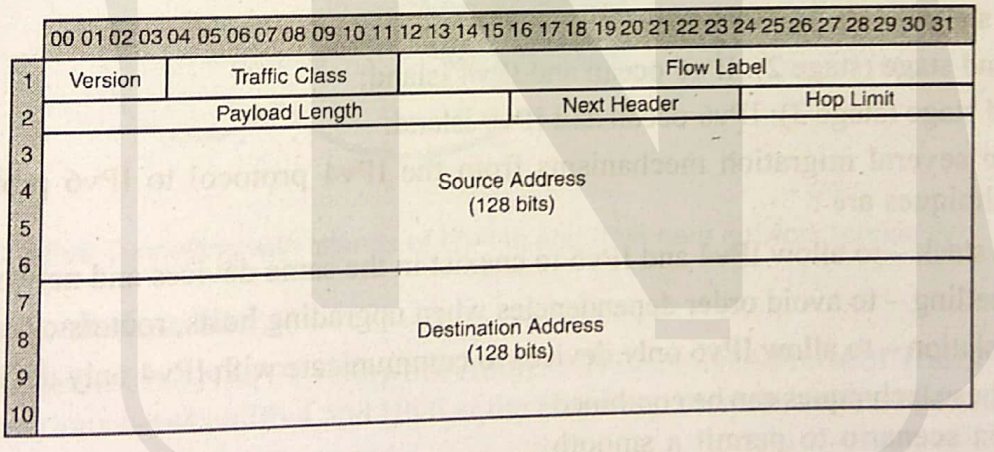

Fig. 7.9. IPv6 Structure

Services and Equipments

The "converging" new generation communication networks are using and planning to use an IP based network infrastructure with multi-functional end-devices, always on, always reachable peer-to-peer, with mobility, quality of service and end-to-end security. Even non telecom industries such as music, radio and television will be supported in the IP environment. There are applications that need or will benefit from IPv6 such:

- Mobile broadband IP;
- Mobile IP broadcast;

- Peer to peer VoIP;
- Digital radio;
- iTV and IPTV;
- Grids;
- P2P multiplayer games;
- RFID;
- Control networks;
- Remote manufacturing systems;
- Sensor networks;
- Microsoft (native support of IPv6 in the next version of Windows – Longhorn).

There are also a few technologies that will support the migration to IPv6 like:

- Powerline Communication;
- Wi-Fi;
- Wi-Max;
- ZigBee;
- Unlicensed Mobile Access (UMA).

Migration

The current IP-based network will gradually migrate from IPv4 to IPv6. Signalling interworking will need to be supported between the IPv6 network and the existing IPv4 network. Mapping of signalling between IPv6 and IPv4 is required. From the deployment point of view, there are three stages of evolution scenarios:

- First stage (stage 1): IPv4 ocean and IPv6 island;
- Second stage (stage 2): IPv6 ocean and IPv4 island;
- Third stage (stage 3): IPv6 ocean and IPv6 island.

There are several migration mechanisms from the IPv4 protocol to IPv6 protocol. The most discussed techniques are :

- Dual stack – to allow IPv4 and IPv6 to coexist in the same devices and networks;
- Tunnelling – to avoid order dependencies when upgrading hosts, routers or regions;
- Translation – to allow IPv6 only devices to communicate with IPv4 only devices.

Most of these techniques can be combined in a migration scenario to permit a smooth transition from IPv4 to IPv6. In the following subsections these three techniques are described briefly.

7.7.5 Dual Stack Technique

In this method it is proposed to implement two protocols stacks in the same device. The protocol stack used for each link depends on the device used at the other end of the link. Fig. 7.4 shows this arrangement.

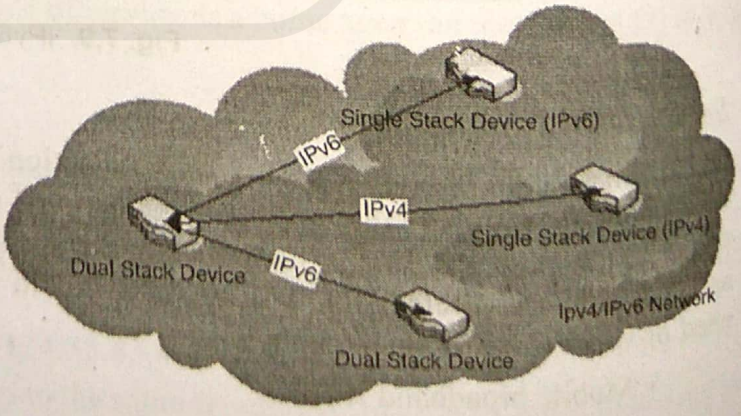

Fig. 7.10. Dual stack operation

7.7.6 Tunnelling Techniques

Tunnelling techniques are used in two phases in the migration to a fully IPv6 network. In the first phase the core of the network uses the IPv4 protocol and there are only small islands IPv6. Figure 5 shows this phase. The IPv6 protocol is encapsulated in IPv4 tunnels.

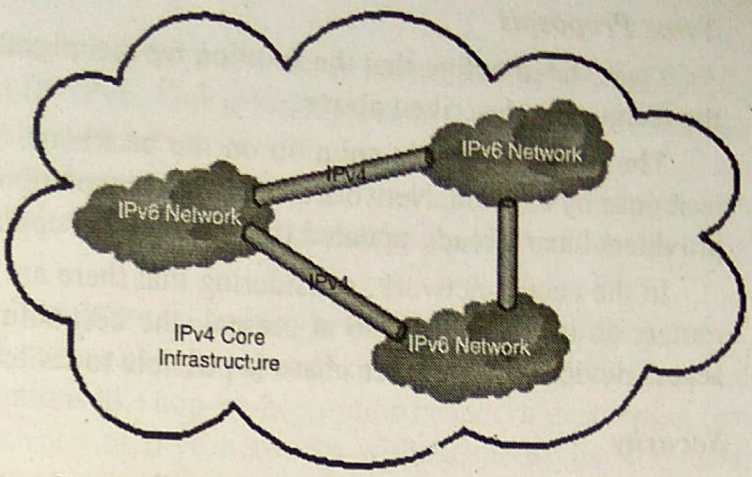

Fig. 7.11. IPv4 Tunnelling with islands of IPv6 in and IPv4 core network (phase 1)

In a second phase, when many nodes in the core of the network have already changed to IPv6, the situation is reversed and

IPv4 is encapsulated in IPv6 tunnels. The following figure shows this second phase.

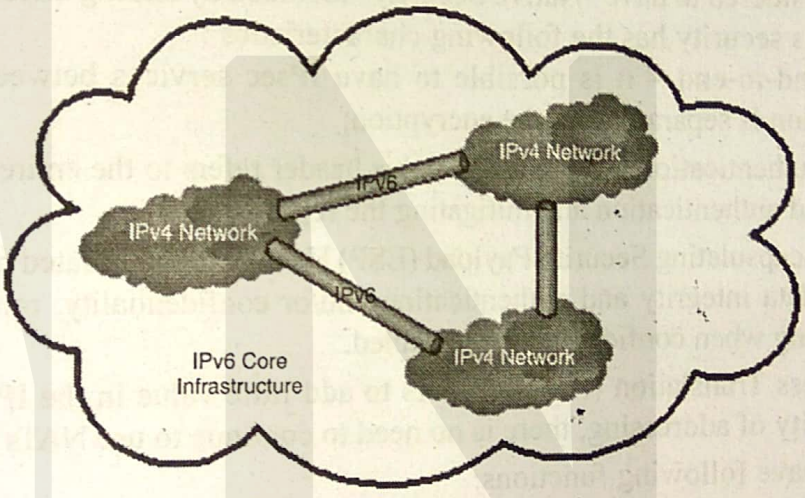

Fig. 7.12. IPv6 Tunnelling with islands of IPv4 in and IPv6 core network (phase 2)

7.7.7 Translation Techniques

This technique uses a device, the NATPT (Network Address Translation – Protocol Translation) that translates in both directions between IPv4 and IPv6 at the boundary between an IPv4 network and an IPv6 network. Figure 7 shows this arrangement.

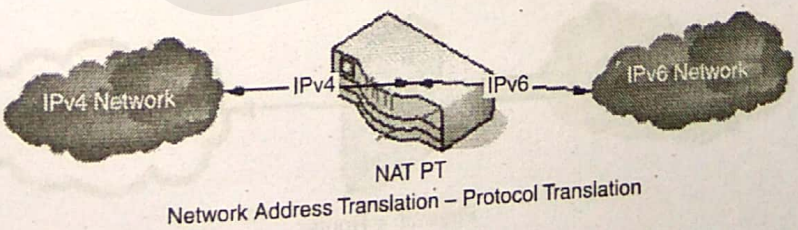

Fig. 7.13: The arrangement with Network Address Translation – Protocol Translation

Some Proposals

As it was stated before that the solution for the migration from IPv4 to IPv6 will be a combination of the techniques described above.

The most preferable solution on the backbone of the IP Network is the use of the dual stack technique by ISPs and Network operators. This solution is already possible, because almost all hardware providers have already updated the software to support this technique.

In the access network, considering that there are many routers at the user's premises and these routers do not support IPv6 at present, the best solution is the use of the translation technique by access devices. In a further phase is possible to switch to the dual stack technique.

Security

"Security is the most common concern with regard to the Internet and to financial transactions via the Internet in particular. Security issues such as authenticating users, controlling access to resources, encrypting communications, and generally ensuring the privacy of transactions all need to be addressed" (European Commission).

The IPv6 is considered to have "Native Security" included by adding different extensions headers in the protocol. This security has the following characteristics :

- It works end-to-end – it is possible to have IPsec services between a pair of hosts; the authentication is separate from the encryption;
- It has an Authentication Header (AH) – this header refers to the entire packet; providing data integrity and authentication and mitigating the replay;
- It has an Encapsulating Security Payload (ESP) Header - encapsulated payload packet (tunnel); providing data integrity and authentication and/or confidentiality; mitigating the replay and limits sniffing when confidentiality is enabled.

Network Address Translation (NAT) appears to add little value in the IPv6 environment. With the increased capacity of addressing, there is no need to continue to use NATs to conserve addresses.

The Firewalls have following functions:

- They enforce uniform policy at perimeter;
- They stop outsiders from performing dangerous operations;
- They provide a check point and scalable, centralised control.

In an IPv6 netwo k end-to-end connectivity, tunnelling and encryptions can conflict with this policy. To avoid these limitations, in an IPv6 network it is necessary to combine the firewall functions and the router functions in the same equipment and to locate it in the edge of the private network. See figure 8 below:

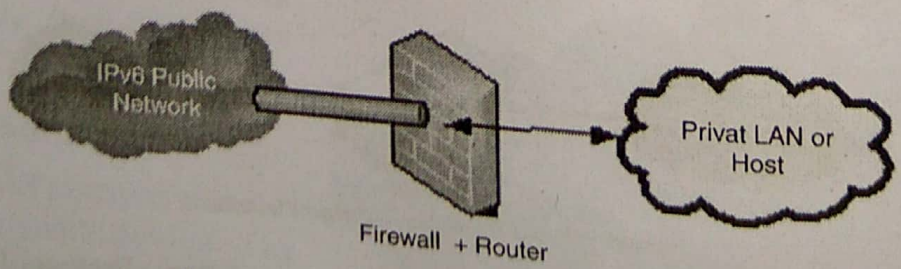

Fig. 7.14. Combined firewall and router

IPv6 and the NGN

The current IP network is in a process of transition from IPv4 to IPv6. Mobile access networks are one of the major potential application areas for IPv6. This is mainly due to the large address space of IPv6. Besides, a large percentile of packets in such networks will carry real time traffic such as voice or video. These applications are expected to be important for IPv6, as they may depend heavily on the QoS mechanism in IPv6 networks. Nevertheless 3GPP is considering using IPv4 addresses initially instead of IPv6 addresses.

From the signalling point of view, the IPv6 protocol has many features related to QoS and other capabilities. By utilising IPv6 features, such as ease of defining explicit route, flow labelling capability and improved support for extensions and options like hop-by-hop option header or destination option header, it is possible to improve the efficiency of IPv6 networks without modifying the existing signalling protocols.

■■■

CHAPTER 8

Latest Concepts and Applications

8.1 CLOUD COMPUTING

Definition

Cloud Computing is a buzzword of 2010 and many experts disagree on its exact definition. But the most used one and concurred one includes the notion of web based services which are available on demand from and optimized and highly scalable service provider. Since such a disagreement on the definition, one will be provided to better understand of the notion. The cloud is IT as a service, delivered by IT resources that are independent of location. It is a style of computing in which dynamically scalable and often virtualized resources are provided as a service over the Internet where end users have no knowledge of, expertise in, or control over the technology infrastructure (the cloud) that supports them.

Attributes

Before some of the attributes will be defined, the term cloud should be explained. A cloud has been long used in IT, in network diagrams respectively, to represent a sort of black box where the interfaces are well known but the internal routing and processing is not visible to the network users. Key attributes in cloud computing:

- **Service Based:** Consumer concerns are abstracted from provider concerns through service interfaces that are well defined. The interfaces hide the implementation details and enable a completely automated response by the service provider. The service could be considered "ready to use" or "off the shelf" because it is designed to serve the specific needs of a set of consumers, and the technologies are tailored to that need rather than the service being tailored to how the technology works. The articulation of the service feature is based on service levels and IT outcomes such as availability, response time, performance versus price, and clear and predefined operational processes, rather than technology and its capabilities. In other words, what the service needs to do is more important than how the technologies are used to implement the solution.
- **Scalable and Elastic:** The service can scale capacity up or down as the consumer demands at the speed of full automation (from seconds for some services to hours for others). Elasticity is a trait of shared pools of resources. Scalability is a feature of the underlying infrastructure and

software platforms. Elasticity is associated with not only scale but also an economic model that enables scaling in both directions in an automated fashion. This means that services scale on demand to add or remove resources as needed.

- **Shared:** Services share a pool of resources to build economies of scale and IT resources are used with maximum efficiency. The underlying infrastructure, software or platforms are shared among the consumers of the service (usually unknown to the consumers). This enables unused resources to serve multiple needs for multiple consumers, all working at the same time.
- **Metered by Use**: Services are tracked with usage metrics to enable multiple payment models. The service provider has a usage accounting model for measuring the use of the services, which could then be used to create different pricing plans and models. These may include pay as you go plans, subscriptions, fixed plans and even free plans. The implied payment plans will be based on usage, not on the cost of the equipment. These plans are based on the amount of the service used by the consumers, which may be in terms of hours, data transfers or other use based attributes delivered.
- **Uses Internet Technologies**: The service is delivered using Internet identifiers, formats and protocols, such as URLs, HTTP, IP and representational state transfer Web oriented architecture. Many examples of Web technology exist as the foundation for Internet based services. Google's Gmail, Amazon.com's book buying, eBay's auctions sharing all exhibit the use of Internet and Web technologies and protocols.

History

History of Cloud Computing surprisingly began almost 50 years ago. The father of this idea is considered to be John McCarthy, a professor at MIT University in US, who first in 1961 presented the idea of sharing the same computer technology as being the same as for example sharing electricity. Electrical power needs many households/firms that possess a variety of electrical appliances but do not possess power plant.

Since that time, Cloud computing has evolved through a number of phases which include grid and utility computing, application service provision (ASP), and Software as a Service (SaaS). One of the first milestones was the arrival of Salesforce.com in 1999, which pioneered the concept of delivering enterprise applications via a simple website. The next development was Amazon Web Services in 2002, which provided a suite of cloud based services including storage, computation and even human intelligence. Another big milestone came in 2009 as Google and others started to offer browser based enterprise applications, though services such as Google Apps.

Architecture

A basis infromation about the architecture is provided in this chapter, together with the explanations of relevant terms such as virtualization, Frond/Back end or Middleware.

- Virtualization is best described as essentially designating one computer to do the job of multiple computers by sharing the resources of that single computer across multiple environments.

The Cloud Computing architecture can be divided into two sections, the front end and the back end, connected together through a network, usually Internet. The **Front End** includes the client's computer and the application required to access the cloud computing system. Not all cloud computing systems have the same user interface. Services like Web based e mail programs leverage existing Web browsers like Internet Explorer or Firefox. Other systems have unique applications that provide network access to clients.

The **Back End** of the system is represented by various computers, servers and data storage systems that create the "cloud" of computing services. Practically, Cloud Computing system could include any program, from data processing to video games and each application will have its own server.

A central server administers the system, monitoring traffic and client demands to ensure everything runs smoothly. It follows a set of rules called protocols and uses a special kind of software called **Middleware**. Middleware allows networked computers to communicate with each other.

Public Cloud (external cloud) is a model where services are available from a provider over the Internet, such as applications and storage. There are free Public Cloud Services available, as well as pay per usage or other monetized models. **Private Cloud** (Interna' Cloud/Corporate Cloud) is computing architecture providing hosted services to a limited number of people behind a company's protective firewall and it sometimes attracts criticism as firms still have to buy, build, and manage some resources and thus do not benefit from lower up front capital costs and less hands on management, the core concept of Cloud Computing.

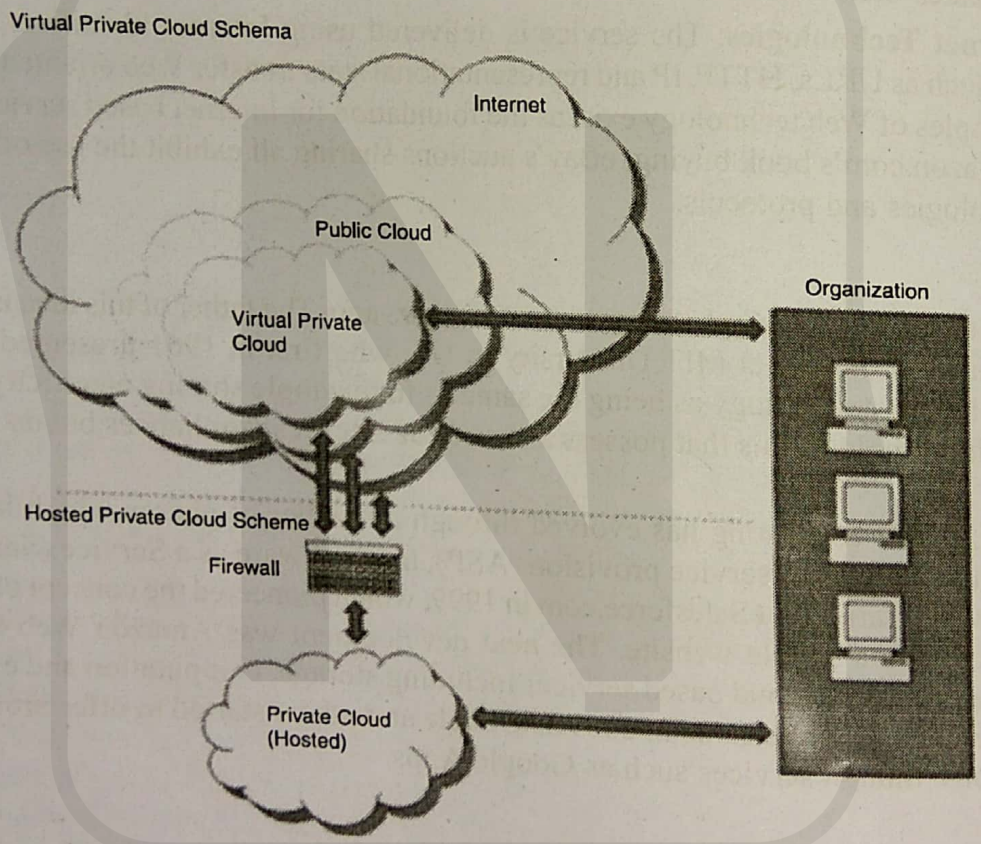

Fig. 8.1. Private/Public cloud

Intergration

Once the definition, categories and componencts needed for the user's solution have been identified the next challenge is to determine how to put them all together. This chapter provides information about the Cloud Computing degisn and integrability as well as gives some examples.

End to end design - definition

It is a major feature of the Internet. The intelligence and functions in an Internet based application reside at both ends of the network (client side and server side), not within the Internet backbone. The Internet acts as a transport between these two.

- **Technical design** – in its simplest form, the end to end design will include the end user device, user connectivity, Internet, cloud connectivity, and the cloud itself.

At a minimum, most organizations will have users who connect to the cloud service remotely (from home or while travelling) and through the internal network. In addition to connectivity at the network level, the interfaces at the application layer need to be compatible and it will be necessary to ensure this connectivity is reliable and secure.

- **Devices** – cloud services should be device agnostic. They should work with traditional desktop, mobile devices and thin client. Unfortunately, this is much easier said than done. Regression testing on five or ten client platforms can be challenging. A good start is to bundle the sets of supported devices into separate services. With Microsoft Exchange 2007 you have the option of supporting Windows platforms through HTTP (Outlook web access) and using RPC over HTTP. You can also support Windows Mobile (as well as Symbian, iPhone and Blackberry devices using ActiveSync). The platform is just beginning. You would also want to take an inventory of existing systems to determine the actual operating platforms, which might range from Mac OS and Linux to Google Chrome, Android, Symbian, RIM Blackberry and iPhones.
- **Connectivity** – in order to assess the connectivity demands you need to identify all required connections. At high level the connections will include categories such as:
  - Enterprise to cloud
  - Remote to cloud
  - Remote to enterprise
  - Cloud to cloud
  - Cloud to enterprise

Once you put these together into a high level connectivity diagram you can then proceed to the next step of identifying and selecting connectivity options. Unless the systems are connected they cannot operate, at least for any extended periods of time. It the case of cloud computing, data and processing are both highly distributed making reliable, efficient and secure connectivity and are the most critical.

- **Management** – generally, for each component in the design we need to investigate how we will manage it. This includes all the end user devices, the connectivity, and legacy infrastructure and all the applications involved. The challenge of splitting management components will be that you may have policies that need to be kept synchronized. Imagine for example, that you have a minimum password length of 8 characters which is increased to 10. If you have only two management servers and this is not a frequent type of occurrence then you can easily apply challenge manually. However, if you are dealing with hundreds of management servers and you receive minor policy changes on a weekly basis you can imagine how cumbersome and error prone the task will become.
- **Security** – the impact of Cloud Computing on security is profound. There are some benefits and unfortunately some hurdles to overcome. One challenge in trying to evaluate security is that it tends to relate to all aspects of IT and, since Cloud Computing's impact is similarly pervasive. Security domains:
- ***Access control*** – provides mechanism to protect critical resources from unauthorized access and modification while facilitating access to authorized users

- ***Cryptography*** presents various methods for taking legible, readable data, and transforming it into unreadable data for the purpose of secure transmission, and then using a key to transform it back into readable data when it reaches its destination.
- ***Operations security*** – includes procedures for back ups and change control management.

The **Cloud Computing Manifesto** is a manifesto containing a "public declaration of principles and intentions" for cloud computing providers and vendors, annotated as "a call to action for the worldwide cloud community" and "dedicated belief that the cloud should be open".

Examples

Most common public known examples of a Cloud are **Google Apps**. This service provide number of on line applications like Word processor, Application for creating and editing presentations, documents storage and sharing, email functions with connection on MS Outlook or MS exchange services, account and contacts sharing, Instant Messenger functions, etc., all provided by Google. Other Clouds examples include CloudX Technology Group, Yahoo, E bay, Facebook, Citric XennApp, AJAX, etc.

Device using CC

- **Chromebook** is a mobile device running Google Chrome OS. The two first devices for sale are by Samsung and Acer Inc. and are slated for release on June 15, 2011 [14] Chromebook (CR 48) is Google prototype model. These machines boot up very quickly and offer basic tools for internet communication. Such as 3G/4G and Wifi connectivity, Web cam and microphone, mobile processor and enought RAM for webbrowsing and works on line only. Basic Hardrive is optional.

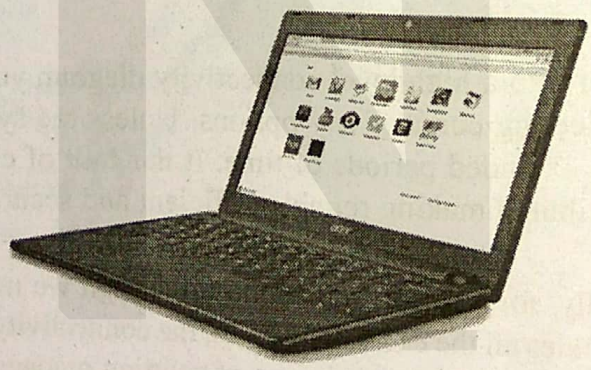

Fig. 8.2. Chromebook by Acer

Pros and Cons

Neither Cloud Computing is an exception and experience both prons and cons. Some of them are stated and described in more details in this chapter.

Pros

- **Lower costs** the principle of sharing resources (HW, SW, infrastructure...) gives to customer also the benefit of sharing its costs. Customer do not has to buy expensive hardware, such as powerful workstations, large server solution and software applications. Customer needs only internet connection and basic PC with not high requirements. Simple laptop, netbook or mobile phone is enought. Customer also pays only for what the real usege. These could be services, hardware resources or infrastructure or its combination.

- **Less IT employees** - there is also no neccesary by customer to employ IT department in such wide range. There is only need to provide secure connection and PC with webrowser. For all other, the technical support such as back-ups, recovery, virus protection, updates, software and hardware stability and functionality, helpdesk and support is maintained by the provider of a service.
- **No special knowledge** - client (customer) also does not need to have a high knowledge about hardware and complex software applications at all. Client just uses a service throught webrowser. Harware resources can be shared between all clients and managed by usage or their requirements.
- **Easy to upgrade** - massive increase of performance (such as speed or storage size) is provided immidiately after simple order and applied by "a few clicks". Data centre can provide higher performance than common desktop PC or, on the other hand, can be very efficient and deliver just what customer needs at the moment (low performance) and thus again it saves resources and money. This approach saves also time, costs for new hardware, transport, is power (energy) efficient and as a result saves the environment, which is very discussed issue these days.
- **Instant access anywhere** - one of the most important benefit is availability of a service anywhere. What is needed for accessing the service is computer connected to the internet. There is no dependence on platform (PC, MAC, mobile phone, car etc.).
- **Security** - is a very discussed issue in the Cloud Computing service providing and could be put in both pros and cons as you see in a while. Service is protected by usage an authorization. Users identify themselves by using an ID (Username) and Password (or also more sophisticated method such as chip, fingerprint, face detection etc. can be used). Communication between client and provider servers is secured. Data centre is protected by firewalls and kept in secured buildings. There generally there is a very low risk of danger caused by attack of third parties. BUT on the other hand, a problem could be that client (customer) keeps all the data out of his computer – just at the providers´ servers. It means the client entrusts the data to the provider (provider company) and has in fact no physical control over them.
- **Requirements** - technology, which customer needs are very simple. Importatnt is only terminal as a laptop, desktop, mobile phone, netbook etc. with web browser, internet connection and usually also created account on a service at providers place.

Cons

- **Legal differences** – as already aforementioned, we can describe one particular example. US companies are obliged to follow the PATRIOT Act (2001) which states that companies can be watched and have to provide information and data about clients, if they are asked for in the correspondence of anti terrorist policy.
- **Dependence on provider** – if company starts using the Cloud Computing service and replaces its previous information system or changes IT structure, it becomes dependant on its service provider. Risks connected with such a dependency may include sudden change of prices or conditions of a contract. Provider could be hit by bankruptcy and end its business activities. Functions and applications might be changed without will of a customer and if a provider suffers from technical problems, all the customers are out of service which means without their data.
- **Reputation** – Cloud Computing is very new type of service. Not many companies has an experience with such a kind of services and application outsourcing. Many users are still worried about data security tranmitted over the internet.

- **Migration costs** – in some cases there can be higher start up costs. Company may have to invest into users training, any amendments which allows the communication of service provider and current company software and in some cases, switching to Cloud Computing could lead to a change of business processes.
- **Less functions** – solutions, which are targeted to the wide range of companies that can't provide specific functions and therefore are not flexible.
- **Dependence on internet connection** all the Cloud Computing applications can be used on line only thus any connection failure could be fatal.

8.2 CLOUD COMPUTING CATEGORIES

There are three main categories in CC, Infrastructure as a Service (IaaS), Software as a Service (SaaS) and Platform as a Service (PaaS). All of them are described below in more details.

- **Infrastructure as a Service** is a provision model in which an organization outsources the equipment used to support operations, including storage, hardware, servers and networking components.
- **Software as a Service** is a software distribution model in which applications are hosted by a vendor or service provider and made available to customers over a network, typically the Internet.
- **Platform as a Service** is an outgrowth of Software as a Service (SaaS). It is a way to rent hardware, operating systems, storage and network capacity over the Internet. The service delivery model allows the customer to rent virtualized servers and associated services for running existing applications or developing and testing new ones.
- **Everything as a Service** (XaaS) is a cloud computing term for the extensive variety of services and applications emerging for users to access on demand over the Internet as opposed to being utilized via on-premises means. Also known as anything -as-a-service facilities the flexibility for users and companies to customize their computing environments to craft the experiences they desire, all on demand. XaaS is dependent on a strong cloud services platform and reliable Internet connectivity to successfully gain traction and acceptance among both individuals and enterprises.

8.3 BIG DATA & DATA ANALYTICS

Definition and Characteristics of Big Data "Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making." — Gartner which was derived from: "While enterprises struggle to consolidate systems and collapse redundant databases to enable greater operational, analytical, and collaborative consistencies, changing economic conditions have made this job more difficult. E-commerce, in particular, has exploded data management challenges along three dimensions: volumes, velocity and variety. In 2001/02, IT organizations much compile a variety of approaches to have at their disposal for dealing each." – Doug Laney

8.3.1 Key Computing Resources for Big Data

- **Processing capability:** CPU, processor, or node.
- Memory
- Storage
- Network

8.3.2 Techniques towards Big Data

- Massive Parallelism
- Data Distribution
- High-Performance Computing
- Data Mining and Analytics
- Machine Learning
- Huge Data Volumes Storage
- High-Speed Networks
- Task and Thread Management
- Data Retrieval
- Data Visualization

8.3.3 More About Big Data & Data Analytics

Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or does not fit the structures of traditional database architectures. In other words, **Big data** is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using on-hand data management tools or traditional data processing applications. To gain value from this data, you must choose an alternative way to process it. Big Data is the next generation of data warehousing and business analytics and is poised to deliver top line revenues cost efficiently for enterprises. Big data is a popular term used to describe the exponential growth and availability of data, both structured and unstructured.

Every day, we create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. This data is **big data**.

Definition

Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, create, manage, and process the data within a tolerable elapsed time

Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision-making.

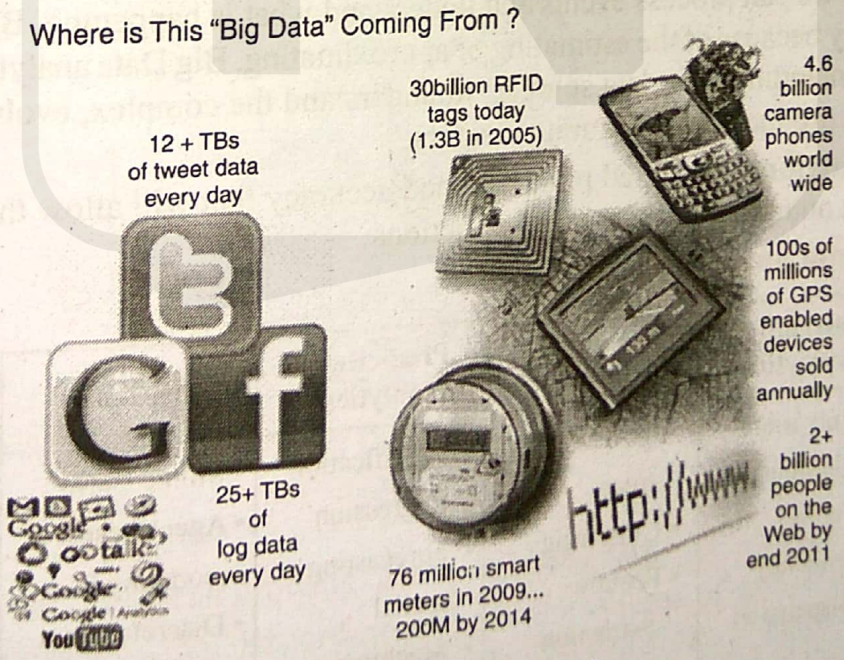

Fig. 8.3. What is Big Data?

Big data is often boiled down to a few varieties including social data, machine data, and transactional data. Social media data is providing remarkable insights to companies on consumer behavior and sentiment that can be integrated with CRM data for analysis, with 230 million tweets posted on Twitter per day, 2.7 billion Likes and comments added to Facebook every day, and 60 hours of video uploaded to YouTube every minute (this is what we mean by velocity of data). Machine data consists of information generated from industrial equipment, real-time data from sensors that track parts and monitor machinery (often also called the Internet of Things), and even web logs that track user behavior online. At arcplan client CERN, the largest particle physics research center in the world, the Large Hadron Collider (LHC) generates 40 terabytes of data every second during experiments. Regarding transactional data, large retailers ar d even B2B companies can generate multitudes of data on a regular basis considering that their transactions consist of one or many items, product IDs, prices, payment information, manufacturer and distributor data, and much more. Major retailers like Amazon.com, which posted $10B in sales in Q3 2011, and restaurants like US pizza chain Domino's, which serves over 1 million customers per day, are generating petabytes of transactional big data. The thing to note is that big data can resemble traditional structured data or unstructured, high frequency information.

8.3.4 Big Data Analytics

Big (and small) Data analytics is the process of examining data—typically of a variety of sources, types, volumes and / or complexities—to uncover hidden patterns, unknown correlations, and other useful information. The intent is to find business insights that were not previously possible or were missed, so that better decisions can be made.

Big Data analytics uses a wide variety of advanced analytics to provide

1. **Deeper insights.** Rather than looking at segments, classifications, regions, groups, or other summary levels you 'll have insights into all the individuals, all the products, all the parts, all the events, all the transactions, etc.
2. **Broader insights.** The world is complex. Operating a business in a global, connected economy is very complex given constantly evolving and changing conditions. As humans, we simplify conditions so we can process events and understand what is happening. But our best-laid plans often go astray because of the estimating or approximating. Big Data analytics takes into account all the data, including new data sources, to understand the complex, evolving, and interrelated conditions to produce more accurate insights.
3. **Frictionless actions.** Increased reliability and accuracy that will allow the deeper and broader insights to be automated into systematic actions.

Advanced Big data analytics

SQL Analytics	Descriptive Analytics	Data Mining	Predictive Analytics	Simulation	Optimization
• Count • Mean • OLAP	• Univariate distribution • Central tendency • Dispension	• Association rules • Clustering • Feature extraction	• Classification • Regression • Forecasting • Spatial • Machine learning • Text analytics	• Monte Carlo • Agent-based modeling • Discrete event modeling	• Linear optimization • Non-Linear optimization

BusinessIntelligence → Advanced Analytics

Big data analytic applications

	Improve Operation Efficiencies	Increase Revenues	Achieve Competitive Differentiation
Mature Analytic Application	• Supply chain optimization • Marketing campaign optimization	• Algorithmic trading	• In-house custom analytic application
Maturing Analytic Application	• Portfolio optimization • Risk management • Next best offer	• Ad targeting optimization • Yield optimization	• In-house custom analytic application
Emerging Analytic Application	• Chronic disease prediction and prevention	• Customer churn prevention	• Product design optimization • Design of experiments optimization • Brand • Product Market Targeting

8.3.5 3 Dimensions/characteristics of Big Data

3Vs (volume, variety and velocity) are three defining properties or dimensions of big data. Volume refers to the amount of data, variety refers to the number of types of data and velocity refers to the speed of data processing.

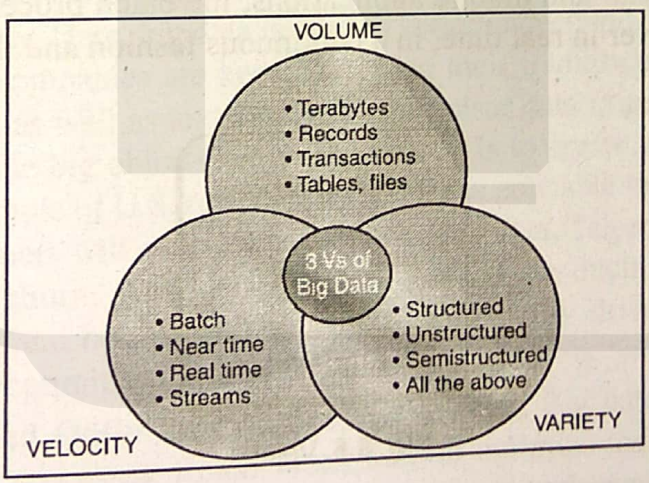

Fig. 8.4. 3 Dimensions of Big Data

Volume

The size of available data has been growing at an increasing rate.

The volume of data is growing. Experts predict that the volume of data in the world will grow to 25 Zettabytes in 2020. That same phenomenon affects every business – their data is growing at the same exponential rate too.

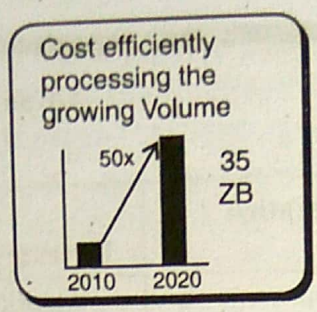

Fig. 8.5. Predicted growth of Big Data from 2010 to 2020

This applies to companies and to individuals. A text file is a few kilo bytes, a sound file is a few mega bytes while a full length movie is a few giga bytes. More sources of data are added on continuous basis. For companies, in the old days, all data was generated internally by employees. Currently, the data is generated by employees, partners and customers. For a group of companies, the data is also generated by machines. For example, Hundreds of millions of smart phones send a variety of information to the network infrastructure. This data did not exist five years ago.

More sources of data with a larger size of data combine to increase the volume of data that has to be analyzed. This is a major issue for those looking to put that data to use instead of letting it just disappear.

Peta byte data sets are common these days and Exa byte is not far away.

Velocity

Data is increasingly accelerating the velocity at which it is created and at which it is integrated. We have moved from batch to a real-time business.

Initially, companies analyzed data using a batch process. One takes a chunk of data, submits a job to the server and waits for delivery of the result. That scheme works when the incoming data rate is slower than the batch-processing rate and when the result is useful despite the delay. With the new sources of data such as social and mobile applications, the batch process breaks down. The data is now streaming into the server in real time, in a continuous fashion and the result is only useful if the delay is very short.

Fig. 8.6. Velocity

Data comes at you at a record or a byte level, not always in bulk. And the demands of the business have increased as well – from an answer next week to an answer in a minute. In addition, the world is becoming more instrumented and interconnected. The volume of data streaming off those instruments is exponentially larger than it was even 2 years ago.

Variety

Variety presents an equally difficult challenge. The growth in data sources has fuelled the growth in data types. In fact, 80% of the world's data is unstructured. Yet most traditional methods apply analytics only to structured information.

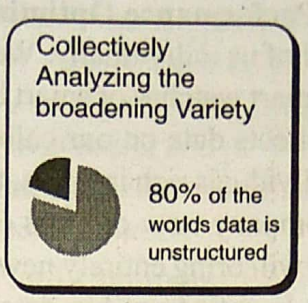

Fig. 8.7: Variety

From excel tables and databases, data structure has changed to loose its structure and to add hundreds of formats. Pure text, photo, audio, video, web, GPS data, sensor data, relational data bases, documents, SMS, pdf, flash, etc. One no longer has control over the input data format. Structure can no longer be imposed like in the past in order to keep control over the analysis. As new applications are introduced new data formats come to life.

The variety of data sources continues to increase. It includes

* Internet data (i.e., click stream, social media, social networking links)
* Primary research (i.e., surveys, experiments, observations)
* Secondary research (i.e., competitive and marketplace data, industry reports, consumer data, business data)
* Location data (i.e., mobile device data, geospatial data)
* Image data (i.e., video, satellite image, surveillance)
* Supply chain data (i.e., EDI, vendor catalogs and pricing, quality information)
* Device data (i.e., sensors, PLCs, RF devices, LIMs, telemetry)

Why Big data?

1. **Understanding and Targeting Customers:** This is one of the biggest and most publicized areas of big data use today. Here, big data is used to better understand customers and their behaviors and preferences. Companies are keen to expand their traditional data sets with social media data, browser logs as well as text analytics and sensor data to get a more complete picture of their customers. The big objective, in many cases, is to create predictive models. You might remember the example of U.S. retailer Target, who is now able to very accurately predict when one of their customers will expect a baby. Using big data, Telecom companies can now better predict customer churn; Wal-Mart can predict what products will sell, and car insurance companies understand how well their customers actually drive. Even government election campaigns can be optimized using big data analytics.
2. **Understanding and Optimizing Business Processes:** Big data is also increasingly used to optimize business processes. Retailers are able to optimize their stock based on predictions generated from social media data, web search trends and weather forecasts. One particular business process that is seeing a lot of big data analytics is supply chain or delivery route optimization. Here, geographic positioning and radio frequency identification sensors are used to track goods or delivery vehicles and optimize routes by integrating live traffic data, etc. HR business processes are also being improved using big data analytics. This includes the optimization of talent acquisition – Moneyball style, as well as the measurement of company culture and staff engagement using big data tools

3. **Personal Quantification and Performance Optimization:** Big data is not just for companies and governments but also for all of us individually. We can now benefit from the data generated from wearable devices such as smart watches or smart bracelets. Take the Up band from Jawbone as an example: the armband collects data on our calorie consumption, activity levels, and our sleep patterns. While it gives individuals rich insights, the real value is in analyzing the collective data. In Jawbone's case, the company now collects 60 years worth of sleep data every night. Analyzing such volumes of data will bring entirely new insights that it can feed back to individual users. The other area where we benefit from big data analytics is finding love - online this is. Most online dating sites apply big data tools and algorithms to find us the most appropriate matches.
4. **Improving Healthcare and Public Health:** The computing power of big data analytics enables us to decode entire DNA strings in minutes and will allow us to find new cures and better understand and predict disease patterns. Just think of what happens when all the individual data from smart watches and wearable devices can be used to apply it to millions of people and their various diseases. The clinical trials of the future won't be limited by small sample sizes but could potentially include everyone! Big data techniques are already being used to monitor babies in a specialist premature and sick baby unit. By recording and analyzing every heart beat and breathing pattern of every baby, the unit was able to develop algorithms that can now predict infections 24 hours before any physical symptoms appear. That way, the team can intervene early and save fragile babies in an environment where every hour counts. What's more, big data analytics allow us to monitor and predict the developments of epidemics and disease outbreaks. Integrating data from medical records with social media analytics enables us to monitor flu outbreaks in real-time, simply by listening to what people are saying, i.e. "Feeling rubbish today - in bed with a cold".
5. **Improving Sports Performance:** Most elite sports have now embraced big data analytics. We have the IBM SlamTracker tool for tennis tournaments; we use video analytics that track the performance of every player in a football or baseball game, and sensor technology in sports equipment such as basket balls or golf clubs allows us to get feedback (via smart phones and cloud servers) on our game and how to improve it. Many elite sports teams also track athletes outside of the sporting environment – using smart technology to track nutrition and sleep, as well as social media conversations to monitor emotional wellbeing.
6. **Improving Science and Research:** Science and research is currently being transformed by the new possibilities big data brings. Take, for example, CERN, the Swiss nuclear physics lab with its Large Hadron Collider, the world's largest and most powerful particle accelerator. Experiments to unlock the secrets of our universe – how it started and works - generate huge amounts of data. The CERN data center has 65,000 processors to analyze its 30 petabytes of data. However, it uses the computing powers of thousands of computers distributed across 150 data centers worldwide to analyze the data. Such computing powers can be leveraged to transform so many other areas of science and research.
7. **Optimizing Machine and Device Performance:** Big data analytics help machines and devices become smarter and more autonomous. For example, big data tools are used to operate Google's self-driving car. The Toyota Prius is fitted with cameras, GPS as well as powerful computers and sensors to safely drive on the road without the intervention of human beings. Big data tools are also used to optimize energy grids using data from smart meters. We can even use big data tools to optimize the performance of computers and data warehouses.

8. **Improving Security and Law Enforcement:** Big data is applied heavily in improving security and enabling law enforcement. I am sure you are aware of the revelations that the National Security Agency (NSA) in the U.S. uses big data analytics to foil terrorist plots (and maybe spy on us). Others use big data techniques to detect and prevent cyber attacks. Police forces use big data tools to catch criminals and even predict criminal activity and credit card companies use big data use it to detect fraudulent transactions.
9. **Improving and Optimizing Cities and Countries:** Big data is used to improve many aspects of our cities and countries. For example, it allows cities to optimize traffic flows based on real time traffic information as well as social media and weather data. A number of cities are currently piloting big data analytics with the aim of turning themselves into Smart Cities, where the transport infrastructure and utility processes are all joined up. Where a bus would wait for a delayed train and where traffic signals predict traffic volumes and operate to minimize jams.
10. **Financial Trading:** My final category of big data application comes from financial trading. High-Frequency Trading (HFT) is an area where big data finds a lot of use today. Here, big data algorithms are used to make trading decisions. Today, the majority of equity trading now takes place via data algorithms that increasingly take into account signals from social media networks and news websites to make, buy and sell decisions in split seconds.

Unstructured data

Unstructured data is information that either does not have a predefined data model and/or does not fit well into a relational database. Unstructured information is typically text heavy, but may contain data such as dates, numbers, and facts as well. The term semi-structured data is used to describe structured data that does not fit into a formal structure of data models. However, semi-structured data does contain tags that separate semantic elements, which includes the capability to enforce hierarchies within the data. The amount of data (all data, everywhere) is doubling every two years. Most new data is unstructured. Specifically, unstructured data represents almost 80 percent of new data, while structured data represents only 20 percent. Unstructured data tends to grow exponentially, unlike structured data, which tends to grow in a more linear fashion. Unstructured data is vastly underutilized.

Mining Unstructured Data

Many organizations believe that their unstructured data stores include information that could help them make better business decisions. Unfortunately, it's often very difficult to analyze unstructured data. To help with the problem, organizations have turned to a number of different software solutions designed to search unstructured data and extract important information. The primary benefit of these tools is the ability to glean actionable information that can help a business succeed in a competitive environment.

Because the volume of unstructured data is growing so rapidly, many enterprises also turn to technological solutions to help them better manage and store their unstructured data. These can include hardware or software solutions that enable them to make the most efficient use of their available storage space.

Unstructured Data and Big Data

As mentioned above, unstructured data is the opposite of structured data. Structured data generally resides in a relational database, and as a result, it is sometimes called "relational data." This type of data can be easily mapped into pre-designed fields. For example, a database designer may set up

fields for phone numbers, zip codes and credit card numbers that accept a certain number of digits. Structured data has been or can be placed in fields like these. By contrast, unstructured data is not relational and doesn't fit into these sorts of pre-defined data models.

In addition to structured and unstructured data, there's also a third category: semi-structured data. Semi-structured data is information that doesn't reside in a relational database but that does have some organizational properties that make it easier to analyze. Examples of semi-structured data might include XML documents and NoSQL databases.

The term "big data" is closely associated with unstructured data. Big data refers to extremely large datasets that are difficult to analyze with traditional tools. Big data can include both structured and unstructured data, but IDC estimates that 90 percent of big data is unstructured data. Many of the tools designed to analyze big data can handle unstructured data.

Implementing Unstructured Data Management

Organizations use of variety of different software tools to help them organize and manage unstructured data. These can include the following:

- **Big data tools**: Software like Hadoop can process stores of both unstructured and structured data that are extremely large, very complex and changing rapidly.
- **Business intelligence software:** Also known as BI, this is a broad category of analytics, data mining, dashboards and reporting tools that help companies make sense of their structured and unstructured data for the purpose of making better business decisions.
- **Data integration tools:** These tools combine data from disparate sources so that they can be viewed or analyzed from a single application. They sometimes include the capability to unify structured and unstructured data.
- **Document management systems:** Also called "enterprise content management systems," a DMS can track, store and share unstructured data that is saved in the form of document files.
- **Information management solutions:** This type of software tracks structured and unstructured enterprise data throughout its lifecycle.
- **Search and indexing tools:** These tools retrieve information from unstructured data files such as documents, Web pages and photos.

8.4 ELEMENTS OF SOCIAL NETWORK

Social Network Analysis or (SNA) is the methodical study of collections of social relationships. These consist of social actors implicitly or explicitly connected to one another. Social network analysts portray the world to be composed of *entities* (e.g., people, organizations, nodes, vertices) that are joined together by *relationships* (e.g., ties, associations, interactions, memberships, links or edges). SNA focuses on relational data about what transpires between entities in contrast to attribute data about individuals. Network analysts focus on the patterns generated within collections of many connections. For individuals, SNA is more about "who you know" than "what you know" or "who you are." At the group level, SNA illuminates how each person's individual connections aggregate to form emergent macro-structures like densely connected sub-groups. Using the mathematics of graph theory, social network analysts calculate and visualize the properties of networks and the social actors that inhabit them. SNA is thus, a method of analyzing data, more than it is a theoretical framework, though there are many theories that have been developed that focus on network data.

8.4.1 Types of Social Networks

There are many types of networks. The specific type of network will determine how to appropriately analyze, visualize, and interpret the data. The type of network is determined by the underlying phenomena it represents. For example, a network of Twitter Following relationships is different from a network of Facebook Friendships because Facebook friendships must be mutual (if you are my Friend I am necessarily your Friend), while Twitter follow relationships don't have to be mutual (I can Follow you without you Following me).

Below is a brief description of the key terminology used to characterize networks.

- **Directed versus Undirected.** Directed networks represent phenomena where the connection between two nodes is not necessarily reciprocated. Undirected networks are always mutual. For example, friendship networks (such as on Facebook where one can't friend another person without their consent) and affiliation networks (e.g., we are connected because we are affiliated with the same organization or we both edit the same wiki page).
- **Weighted versus Unweighted.** Some edges have values associated with them. For example, edges in an email network are "weighted" based on the number of messages one person sends to another person, while a wiki co-edit page network is weighted based on the number of pages two people have both edited. Other edges are binary; they either exist or they don't. For example, Facebook friendships and Twitter follow relationships don't have weights.
- **Multiplex Networks.** Multiplex networks include multiple types of edges. For example, a network that connects people together based on their communication via email, phone, and face-to-face interactions would include 3 distinct types of edges. This could be analyzed and visualized as a single multiplex network or as 3 distinct networks.
- **Unimodal and Multimodal Networks**. Many social networks, called unimodal networks, include only one type of node. For example, all the nodes represent people. Or, all of the nodes represent organizations. In contrast, multimodal networks include more than one type of node. For example, a network may include people who are connected to organizations, or another network may include people who are connected to wiki pages they have edited. If there are only two types of nodes we call the network bimodal, which is a subset of the more general multimodal concept. Many bimodal networks, called bipartite networks, have one type of node (i.e., people) connected to another type of node (e.g., organizations) without any edges connecting nodes of the same type (e.g., people to people). These bipartite networks can be transformed into unimodal networks. For example, the person-to-organization network can be transformed into a person-to-person network where people are connected by a weighted edge that represents the number of organizations they are both a part of. Conversely, an organization-to-organization network could be created where the a weighted edge represents the number of people who are part of both organizations.
- **Partial Networks**. In practice, it is not practical or useful to collect data on an entire network (e.g., all Facebook users). Instead, analysts create partial networks in a variety of ways. One approach is to create an "egocentric network", which includes a single node (called "ego") and all of the nodes that ego is directly connected to (called "alters"). When the connections between alters are also included, the graph is called a 1.5 degree network. Adding ego's "friends of friends" makes it a 2.0 degree network and so forth. Other techniques for creating partial networks include sampling a large network (Leskovec&Faloutsos, 2006) or finding some network boundary such as membership in an organization.

8.4.2 Network Analysis Tools

Social network analysis requires the use of specialized software designed to compute network metrics and visualize network graphs. The tool landscape is in constant flux

Table 3. Commonly Used Network Analysis and Visualization Tools

SNA Tool	Description	Expertise required	Open Source	Maximum Network Size
Gephi	Standalone network analysis designed primarily for visualization. Can be extended via plugins.	Designed for novices.	Yes	hundreds of thousands
NodeXL	Includes sophisticated graph visualizations, social media data importers, and extensibility via formulas and macros, but relatively few metrics.	Microsoft Excel plugin designed for SNA novices.	Yes	tens of thousands
Pajek	Includes comprehensive list of network metrics and statistical tests. Steep learning curve.	Designed for sophisticated analysis of large datasets.	Yes	millions
R	Open source statistical package with social network analysis functionality via the igraph, sna, network, and statnet packages. Includes comprehensive list of network metrics and statistical tests.	Steep learning curve.	Yes	millions
UCINet	Includes comprehensive list of network metrics and statistical tests. Designed for knowledgeable SNA researchers, but does not require coding.	Designed for researchers performing social network analyses.	No	tens of thousands