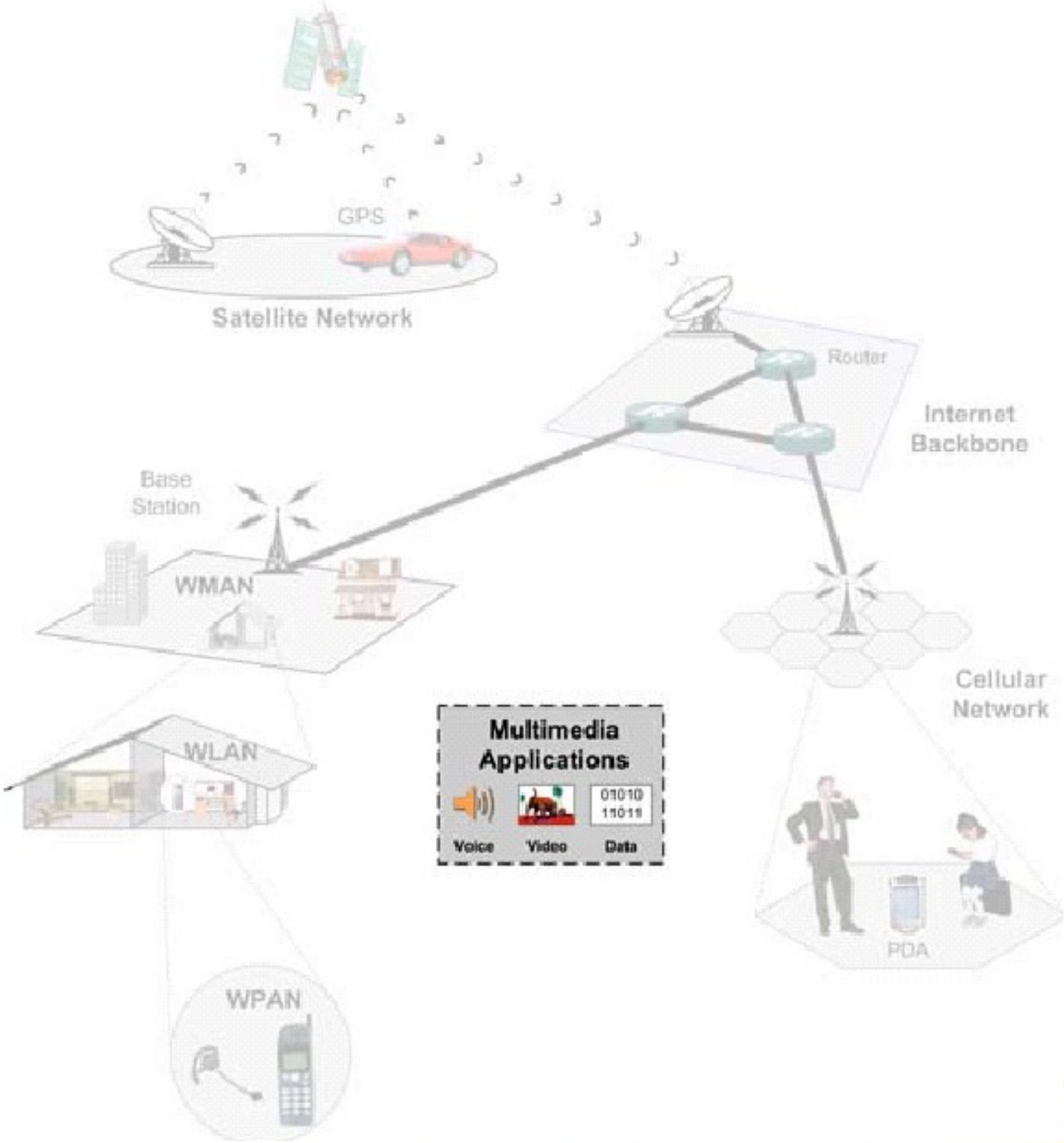
Part-I: Multimedia Applications and Quality of Service (QoS)

鄭瑞光

台灣科技大學電子系


GPS
Satellite Network
Router
Internet
Backbone
Base
Station
WMAN
Cellular
Network
WLAN
Multimedia
Applications
01010
11011
Voice
Video
Data
PDA
WPAN

Outline

- Multimedia applications and quality of service (QoS)
  - Multimedia Applications
  - QoS Fundamentals
  - QoS Mechanisms
    - Classification
    - Channel Access Mechanism
    - Packet Scheduling Mechanisms
    - Traffic Policing Mechanism
    - Resource Reservation Signaling Mechanisms
    - Admission Control
    - QoS Architecture

Multimedia applications and QoSs

- Chapter 1
  - the nature of multimedia applications
  - users' expectations for multimedia applications
- Chapters 2 and 3
  - QoS fundamental concepts
  - QoS mechanisms

Chapter 1
Multimedia Applications

Multimedia Applications

- Streaming video
- Streaming audio
- Collaboration
- One-way and interactive multimedia messaging
- Gaming, including interactive peer-to-peer (p2p) gaming
- Digital money transactions
- MP3 music download
- Video- and audio-supported shopping

Applications

- Long-distance learning, education
- Video and audio conferencing
- File sharing and transfer (pictures, video clips, and text)
- Feeding of real-time news and information (i.e. weather, financial markets, sports, etc.)
- Geographic location services
- Safety services such as Enhanced 911 (E911)
- Gambling
- Entertainment

- With QoS,
  - The applications can improve interactivity, reduce jitter, and provide continuous video and voice experience.
- Without QoS
  - The applications that tailor location-based information for consumers may provide outdated information when latency is high.

Video and Audio Streaming

- Television may be the most well-known form of streaming video.
  - It already feeds wireless multimedia streams into millions of dishes and antennas, connected to TVs and other devices.
- Two major providers of streaming video in US
  - DirecTV (www.directv.com)
  - Dish Networks (www.dishnetworks.com)

Video and Audio Streaming

- To allow efficient streaming,
  - the provider needs to send the data as a steady stream, and
  - the receiver needs to be able to cache excess data in a temporary buffer until used.
- If the data do not arrive fast enough, users will experience interruptions.
- Several competing streaming technologies,
  - RealAudio, RealVideo (www.real.com),
  - Microsoft Media Player (www.microsoft.com),
  - PacketVideo (www.packetvideo.com), and
  - QuickTime (www.apple.com).

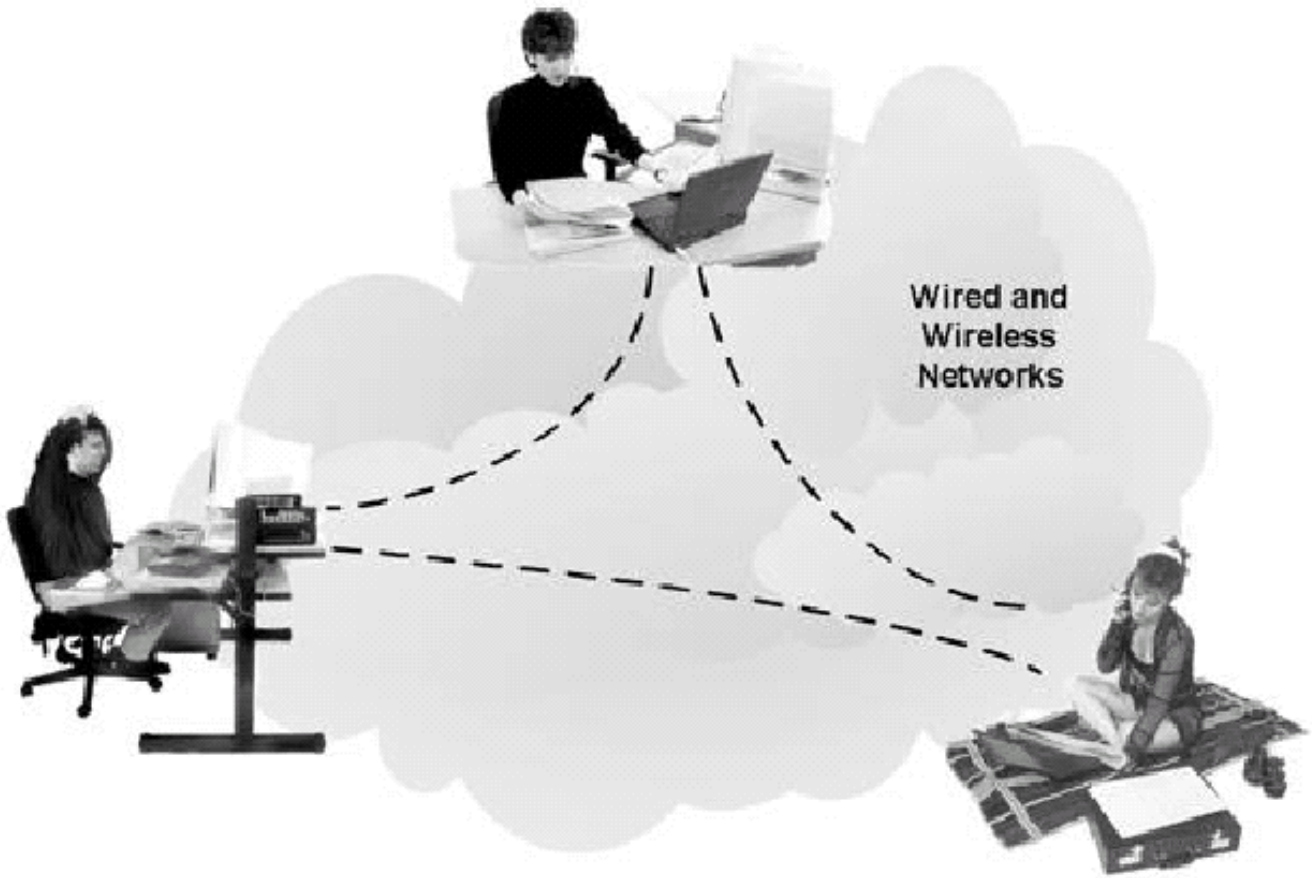
Video and Audio Streaming

- The streaming video quality depends on
  - the capacity of the transmission channel
    - higher data rate is better
  - its ability to support a steady stream
    - steady data rate is better

Peer-to-Peer Computing

- Peer-to-peer (p2p) computing is the sharing of files, memory, computation power, and other computer resources and services among devices.
- p2p Computing aggregates the shared network resources allowing economical execution of applications.
- Examples:
  - Napster started the revolution in music distribution.
  - Groove, Endeavors, and eZmeeting are developing p2p business applications.

Peer-to-Peer Computing

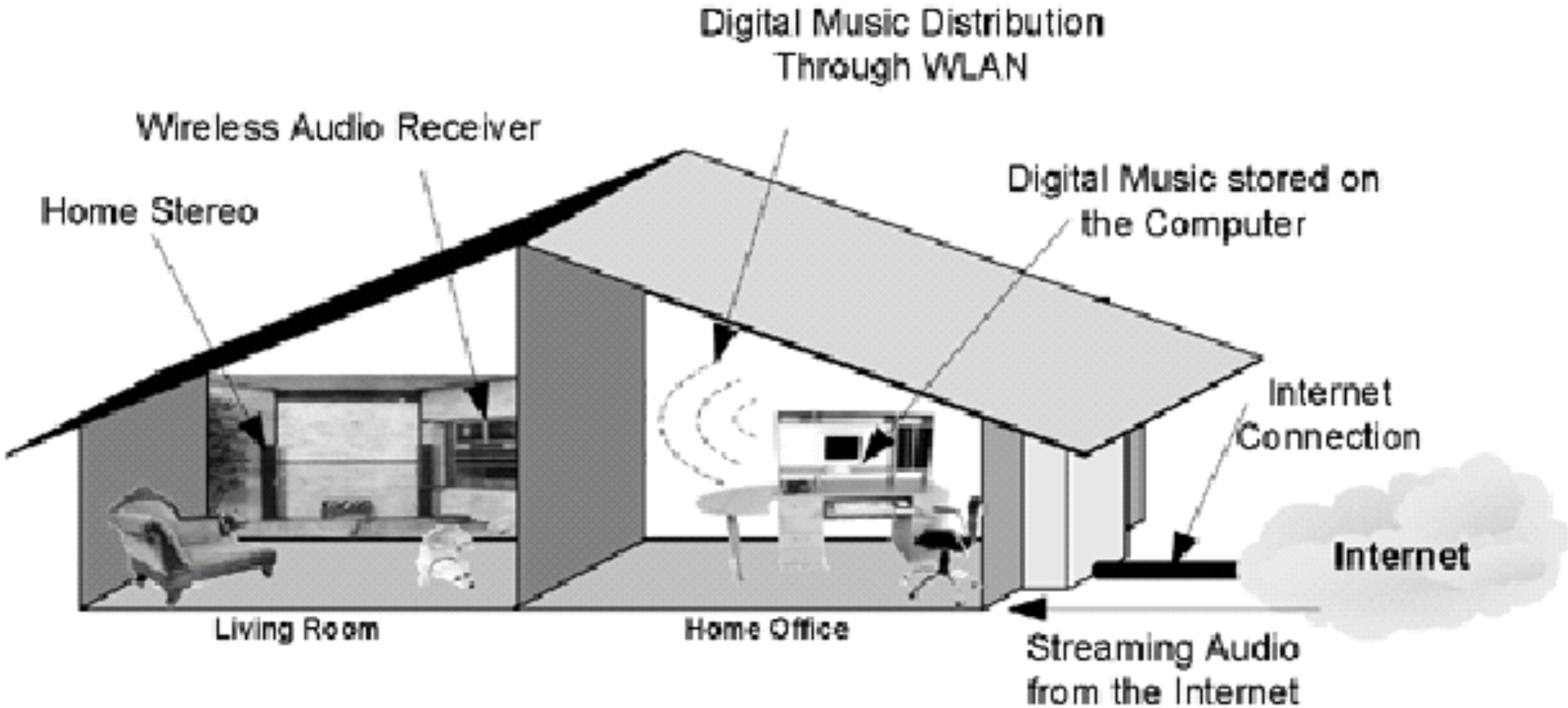


Digital Money Transactions

- Banks will be able to support the consumer who wants to obtain cash from a traditional automated teller machine (ATM) using his or her mobile phone.
- Example
  - Cmode: users can purchase Coca-Cola via their i-mode phones.
  - iConvenience: shopping through i-mode phones.

Entertainment

- One of the most prevalent applications for WLANs is entertainment.



File and Picture Sharing

- Centralized file sharing
  - Online storage services
    - Xdrive (www.xdrive.com),
    - Yahoo! Briefcase (www.briefcase.yahoo.com)
    - StoragePoint (www.storagepoint.net)
- Distributed file sharing
  - Kazaa Media Desktop (www.kazaa.com)
  - Morpheus (www.morpheus.com)

File and Picture Sharing

- Digital pictures
  - Clubphoto (www.clubphoto.com)
  - ofoto (www.ofoto.com)
  - Yahoo! Photos (www.photos.yahoo.com)

Email and Multimedia Messaging

- Protocols:
  - Short Message Service (SMS)
  - Enhanced Messaging Service (EMS)
  - Multimedia Messaging Service (MMS)
- Companies:
  - SpotLife (www.spotlife.com)

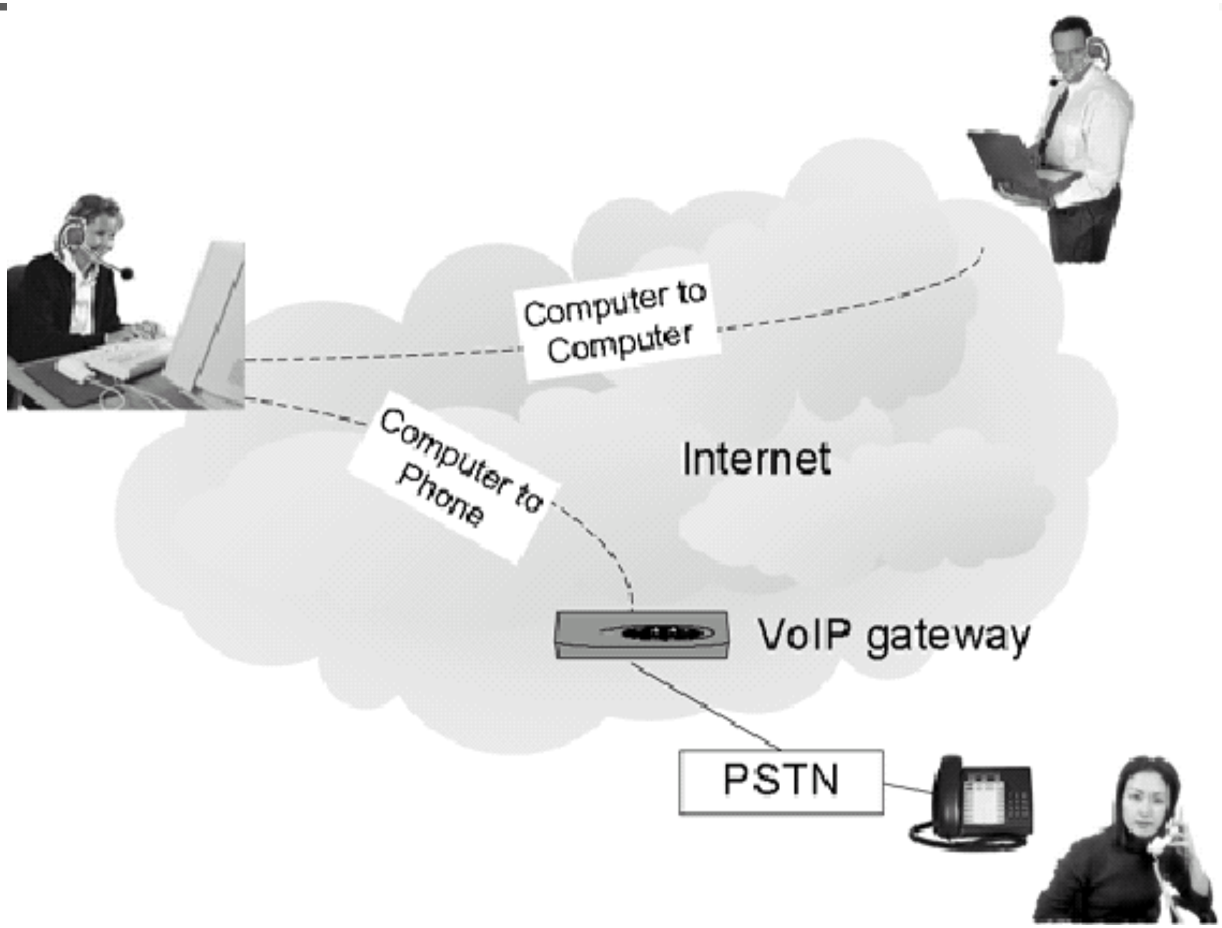
Wireless Gaming

- Gladiator: (古羅馬)鬥士
  - a wireless multiplayer combat game set in the Coliseum (競技場) of ancient Rome.
  - Players choose characters and weapons for their gladiators and then duel against a live opponent in real time.
  - In 2001 JAMDAT Mobile (www.jamdatmobile.com) reported that, in less than three months, 300k mobile device users spent more than 3M min.s
  - Players accessed their proprietary game server from more than 20 wireless carriers around the world using more than 65 different devices.

Voice and Telephony

- Voice and telephony services over wireless networks
  - cellular phones
  - cordless phones
  - VoIP

VoIP

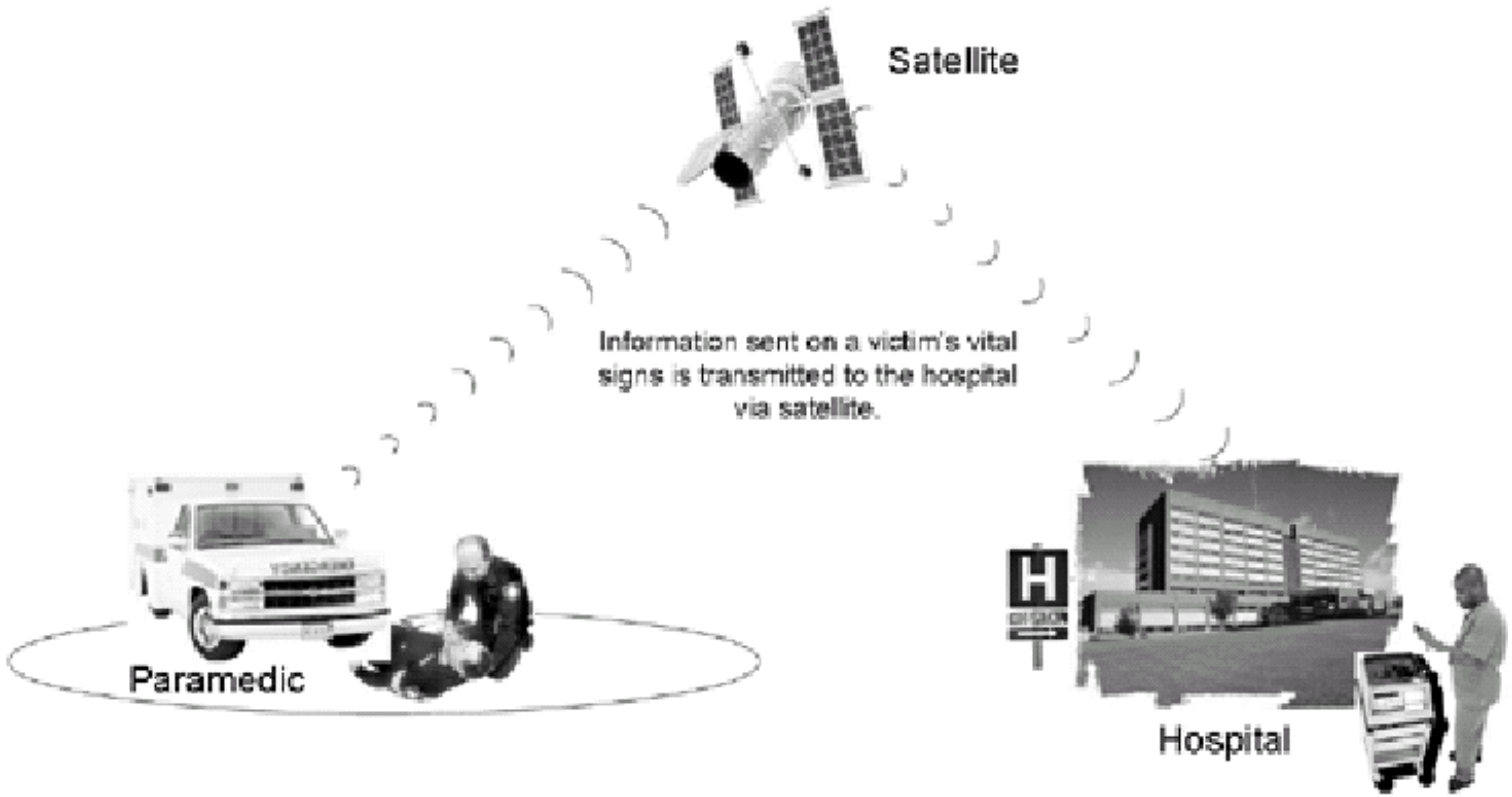

Computer to
Computer
Computer to
Phone
Internet
VoIP gateway
PSTN

Location-Based Services

- Information is sent to a user based on the user's location provided via the GPS (Global Positioning System).
- Applications
  - traffic reporting,
  - restaurant recommendation,
  - navigation
  - customized ads.
  - imagine receiving for shopping

Telemedicine

- Remote Communities Services Telecentre (RCST) project (www.rcst.net)
  - run by a satellite operator
  - provide integrated tele-learning and telemedicine services via broadband satellite links.
    - highspeed Internet access
    - video conferencing
    - digital imaging

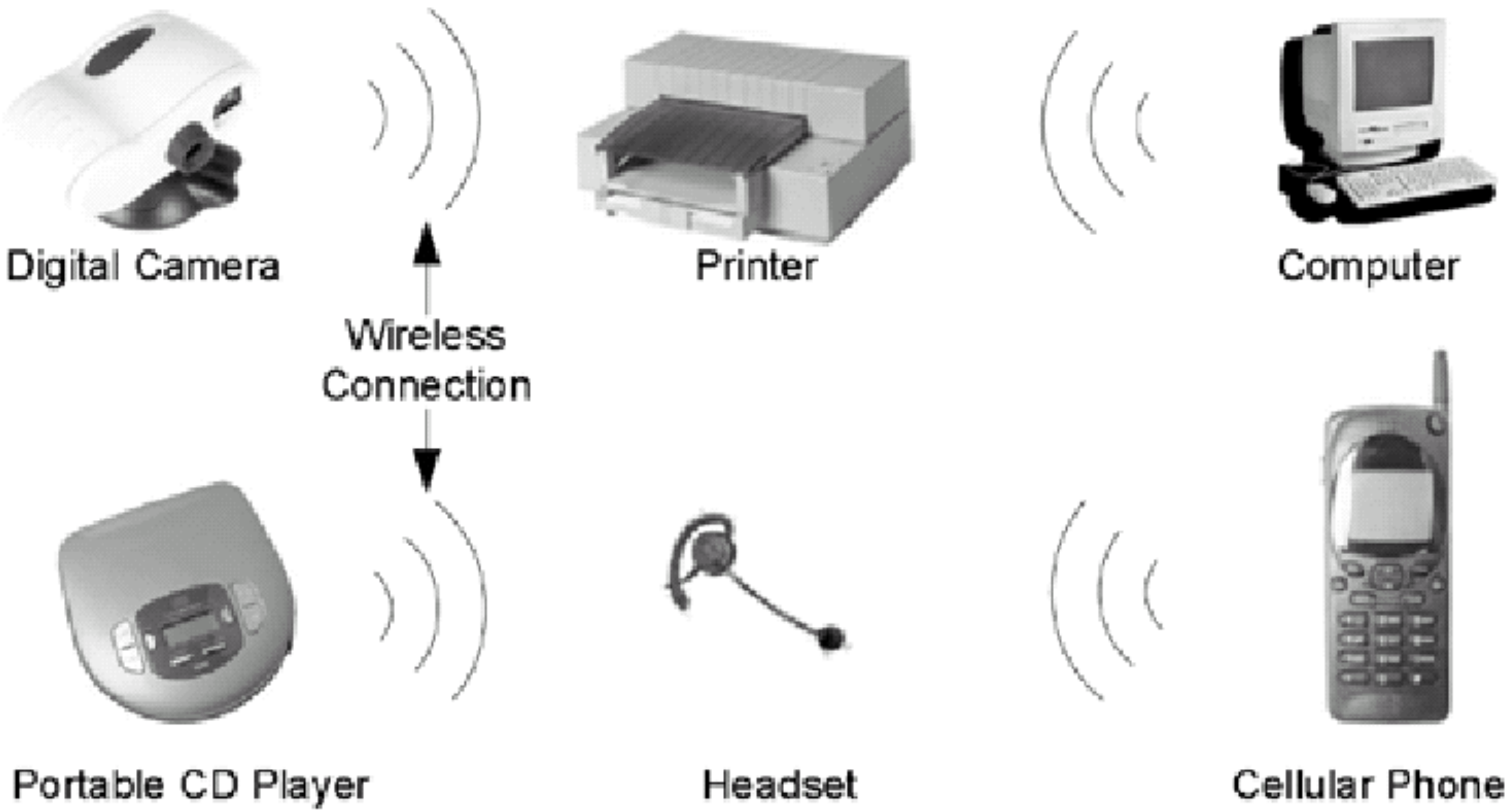

Satellite
Information sent on a victim's vital signs is transmitted to the hospital via satellite.
Paramedic
Hospital

Business Applications

- Business applications
  - payment authorization,
  - distance learning,
  - multicast delivery of promotional content,
  - retail point-of-sale (POS) transaction processing
- Example,
  - General Motors and Ford are able to train its management and technical personnel in any office, anywhere in the world.
  - WalMart, Shell, Texaco, and other companies use this service for their retail POS transactions.

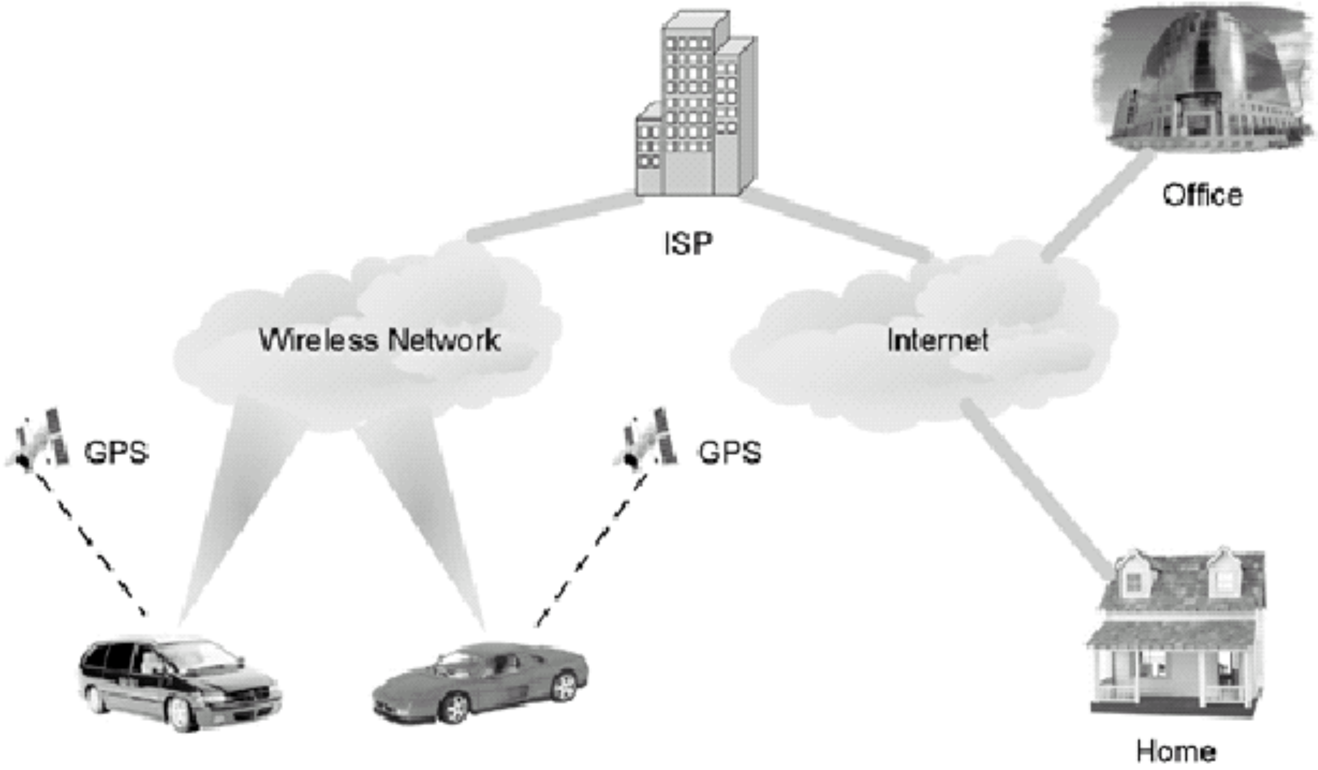
Personal Area Device Connectivity

- Personal area devices refer to devices carried by, placed near, or used near a person.
  - Consumer devices: CD player, digital camera
  - Games: game controller, joystick
  - Professional devices: PDA, notebook
  - Sport training devices: health monitor, sensor
  - Hospital devices: blood pressure sensor
  - Military devices: combat equipment


Digital Camera
Printer
Computer
Wireless Connection
Portable CD Player
Headset
Cellular Phone


Wireless Ear Microphone
Acoustic Antisniping Sensor
Wireless Silent Alarm
Body Core Temperature Sensor
Personal Inertial Navigation Systems
Weight Monitor
Hydration Monitor

Telematics (Automobiles)



Main protocols

- Short Message Service (SMS)
  - text message service
  - Transmits short messages between a cell phone, PDA, PC, or any device with an IP address.
  - Supports messages that are no more than 140 to 160 alphanumeric characters in length
    - ~ 30 to 40 words in Latin
    - ~ 70 characters for non-Latin alphabets like Chinese and Arabic

Main protocols

- Enhanced Messaging Service (EMS)
  - supports pictures, melodies, sound marks, graphic, animations, fonts, and formatted text
  - basic and extended pictures
    - Basic pictures:
      - black and white
        - small (16x16 pixels) pictures
        - large (32x32 pixels) pictures
    - Extended pictures:
      - black and white, grayscale, or color
      - can contain 255x255 pixels

Main protocols

- Enhanced Messaging Service (EMS)
  - Predefined sounds
    - Formatted according to the iMelody standard
    - These melodies can take up to 128 bytes.
  - Animation
    - small (8x8 pixels) animations
    - large (16x16 pixels) animations
    - Extended animations: black & white, grayscale, or color
    - The maximum size of a single animated frame is 255x255 pixels.

Main protocols

- Multimedia Messaging Service (MMS)
  - allowing users to create messages that include any combination of text, graphics, photographic images,speech, and audio or video clips.
  - supports standard
    - image formats: JPEG and GIF
    - video formats: MPEG 4
    - audio formats: MP3 and MIDI
  - No size limitation
  - Use WAP as a supporting technilogy
- Comparison:
  - SMS, EMS: transmitted via control channel
  - MMS: transmitted via data channel

Main protocols

- Wireless Application Protocol (WAP)
  - an open, global standard that provides a microbrowser environment optimized for wireless devices, such as phones and pagers.
  - Goal:
    - to easily access and interact with information over the Internet.
  - WAP 2.0
    - most recent standard published in 2002
    - adds support for IP, TCP, and HTTP
  - WAP uses the Pull Model, where the user requests content from the server

Chapter 2
QoS Fundamentals

Outline

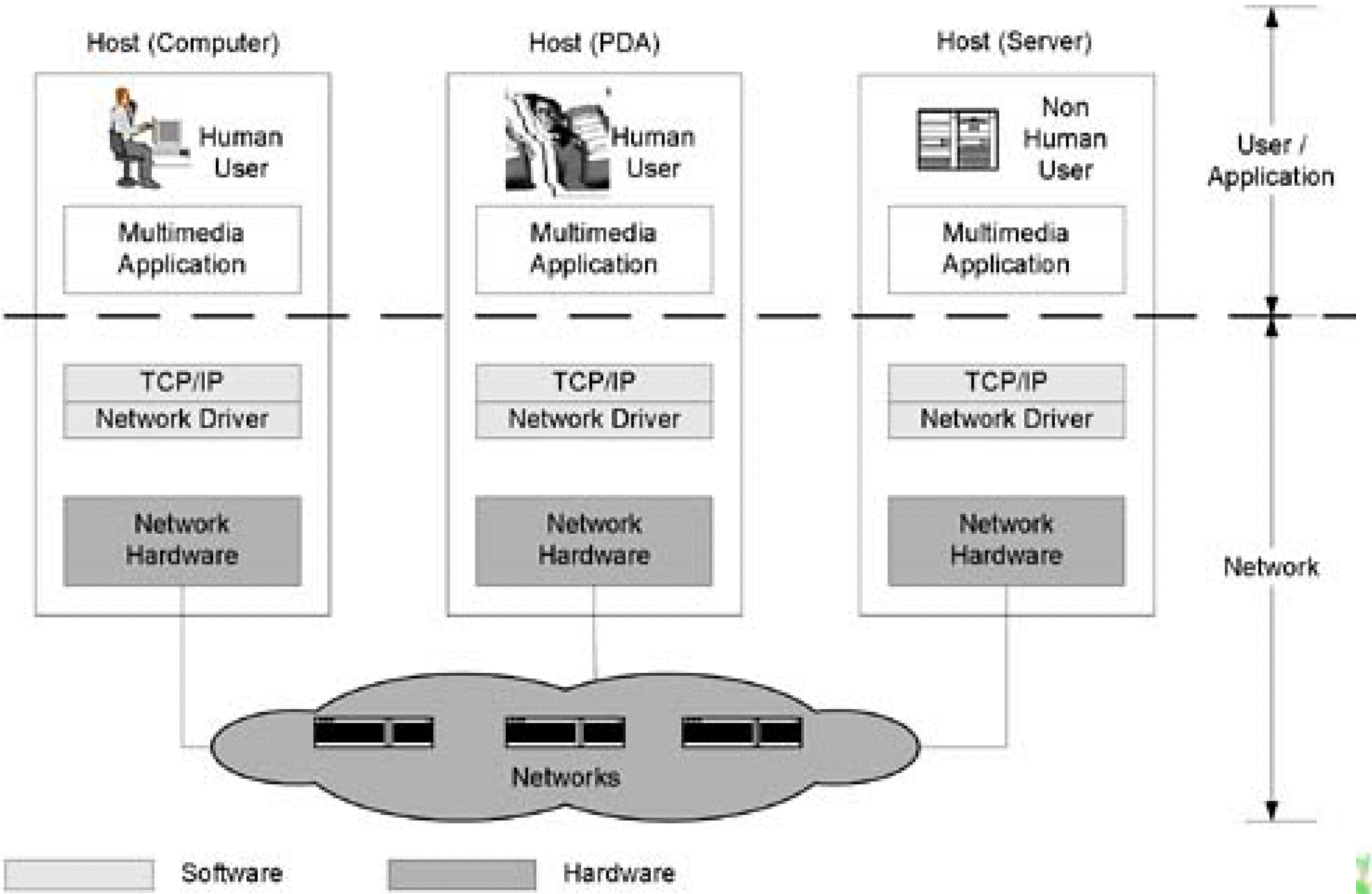
- Introduction
- QoS Parameters
- Multimedia Application Requirements
- QoS Services
- Realization of QoS Services

Introduction to QoS

- Factors and components that affect the performance of multimedia applications
  - *Users:*
    - Human/nonhuman who utilize multimedia applications.
    - Users‘ perception can influence the evaluation of the multimedia applications' performance.
  - *Host machine:*
    - Devices that operate the multimedia application (source and destination hosts).
    - It consist of
      - processors, media storage systems (e.g., hard drive, CD-ROM), display devices, operating systems, etc.

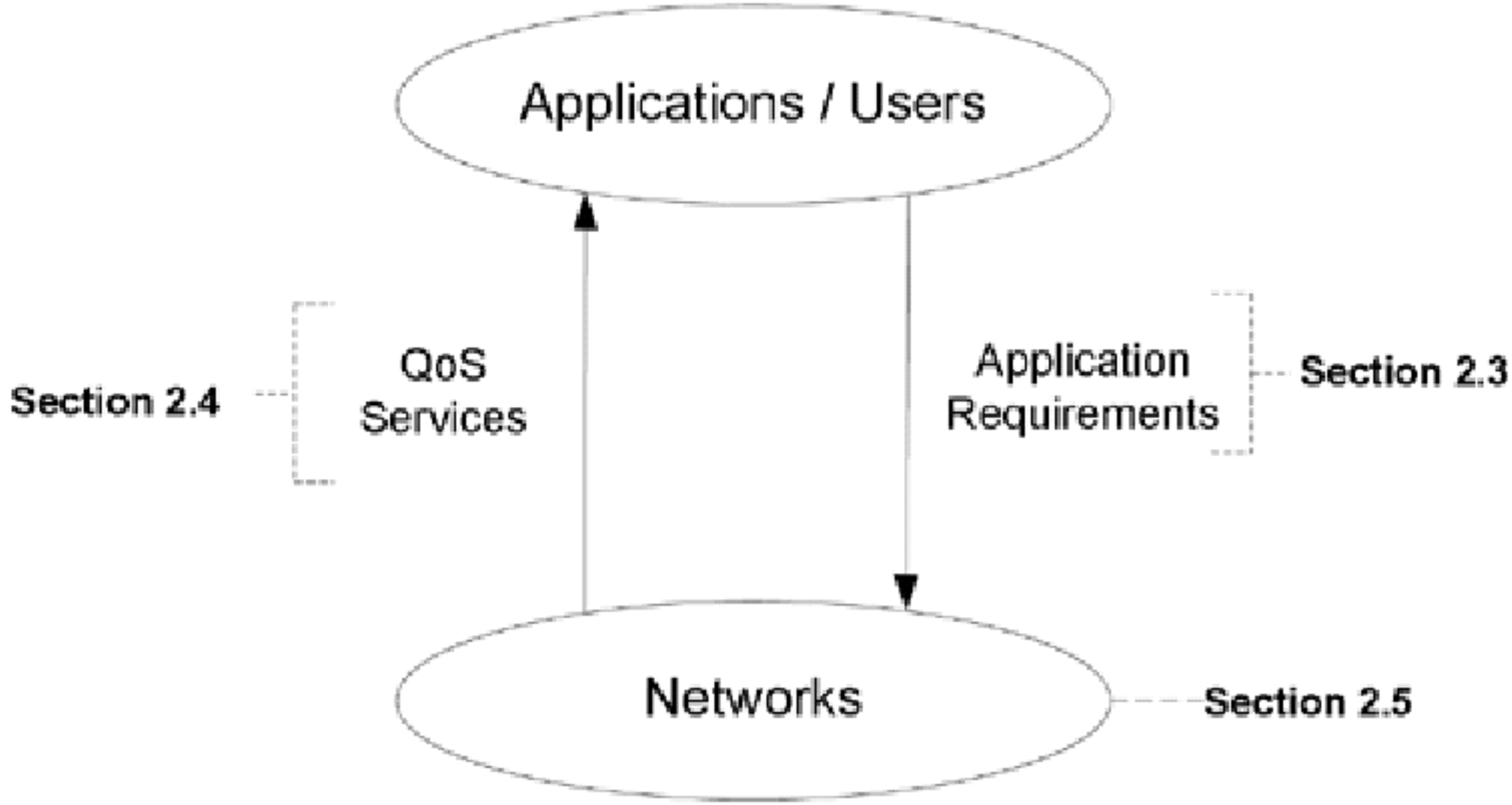
Introduction to QoS

- Factors and components that affect the performance of multimedia applications
  - *Application:*
    - Structure or mechanisms built in the multimedia application (e.g., video codec)
  - *Network:*
    - Network components that transport the multimedia contents between two host machines (source and destination)
    - Examples
      - hardware components: switches, routers, network interface cards, gateways, and firewalls.
      - Software components: network protocols


Host (Computer)
Host (PDA)
Host (Server)
Human User
Human User
Non Human User
Multimedia Application
Multimedia Application
Multimedia Application
TCP/IP
Network Driver
TCP/IP
Network Driver
TCP/IP
Network Driver
Network Hardware
Network Hardware
Network Hardware
Networks
User / Application
Network
Software
Hardware

- Multimedia support issues can be presented by using the Quality of Service (QoS) term.
- Two QoS perspectives:
  - Networks:
    - the service quality or service level that the network offers to applications or users in terms of network QoS parameters.
    - network QoS parameters:
      - latency or delay of packets traveling across the network
      - reliability of packet transmission (i.e. packet lost rate)
      - throughput.
  - Applications/Users
    - the application quality as perceived by the user
    - Example
      - the presentation quality of the video,
      - the responsiveness of interactive voice,
      - the sound quality (CD-like or FM-radio like sound) of streaming audio.

- Users
  - DO NOT concern about how the network manages its resources or what mechanisms are involved in QoS provision.
  - DO concern about the services that networks provide which directly impact the perceived quality of the application.
- Networks
  - Goal:
    - To provide the QoS services that adequately meet the users' needs
    - To maximize the network resources (i.e., bandwidth) utilization.


Applications / Users
Section 2.4
QoS
Services
Application
Requirements
Section 2.3
Networks
Section 2.5

Issues for the QoS model

- What information is contained in the application requirements?
- Desired quality of the application which relates to users' satisfaction is subjective.
  - How does the desired quality map to QoS parameters which are managed by networks?
  - What criteria need to be considered in the mapping process?
- What are the QoS parameters?

Issues for the QoS model

- Which QoS mechanisms do the networks use to achieve diverse QoS support?
- What kind of QoS service level can networks offer?
- What is the relationship between QoS services and application requirements?

QoS Parameters

- QoS parameters which are relevant to multimedia applications:
  - Throughput or bandwidth
  - Delay or latency
  - Delay variation (delay jitter)
  - Loss or error rate
  - Fairness

Throughput

- From the application perspective, throughput refers to the data rate (bits per second) generated by the application.
- Throughput
  - Unit: bits/sec. (bps)
  - Sometimes, it is called *bit rate* or *bandwidth*.
    - Bandwidth is the network resource that needs to be properly managed and allocated to applications.
  - The required throughput depends on the application characteristics.
    - e.g., in a streaming video application, different video properties generate different throughput.

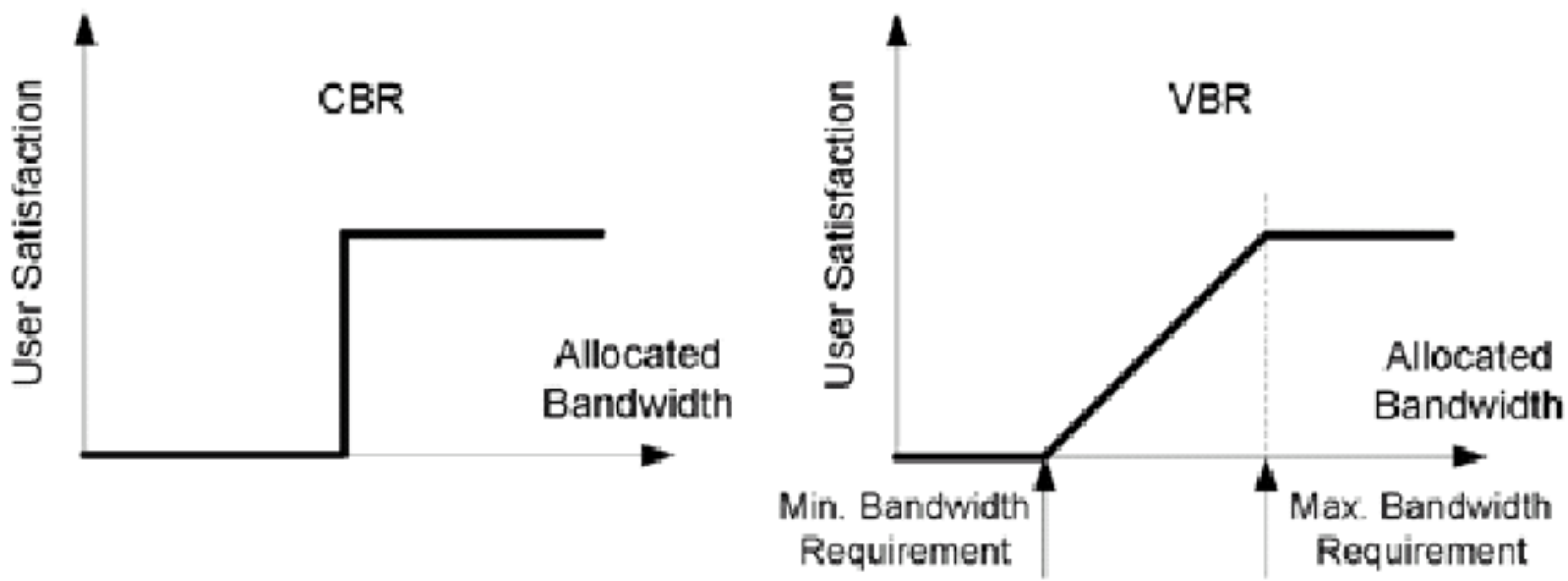
- The video quality can be varied by the following video properties:
  - *Frame size:*
    - Pixels/row, pixels/column, bits/pixel
  - *Frame rate:*
    - Frames/second
    - frame rate ↓ bandwidth consumption ↓ smoothness of the video movement ↓
  - *Color depth:*
    - Colors/pixel
    - 8 bits/pixel→256-color; 24 bits/pixel →16M-color
  - *Compression:*
    - Reduction of the bandwidth consumption at the expense of image quality.
    - Examples: MPEG1, MPEG2, and MPEG4.

Constant and Variable Bit Rate

- *Constant bit rate (CBR) applications*
  - data traffic is generated with constant data rate
  - Example:
    - Digital telephone Private Branch Exchange (PBX), which generates 64 kbps constant bit rate
    - Uncompressed digital video
  - Delay sensitive
  - Bandwidth allocation
    - constant bandwidth allocation

Constant and Variable Bit Rate

- *Variable bit rate (VBR) applications*
  - data traffic is generated with variable data rate
  - Example:
    - Compressed video and audio: delay-sensitive
    - Remote login: non-delay-sensitive
  - Bandwidth allocation
    - minimum bandwidth is required to operate successfully.
    - the more allocated bandwidth, the better the user-perceived quality.
    - user satisfaction is not improved if the maximum required bandwidth exceeds


CBR
User Satisfaction
Allocated Bandwidth
VBR
User Satisfaction
Allocated Bandwidth
Min. Bandwidth Requirement
Max. Bandwidth Requirement

Burstiness

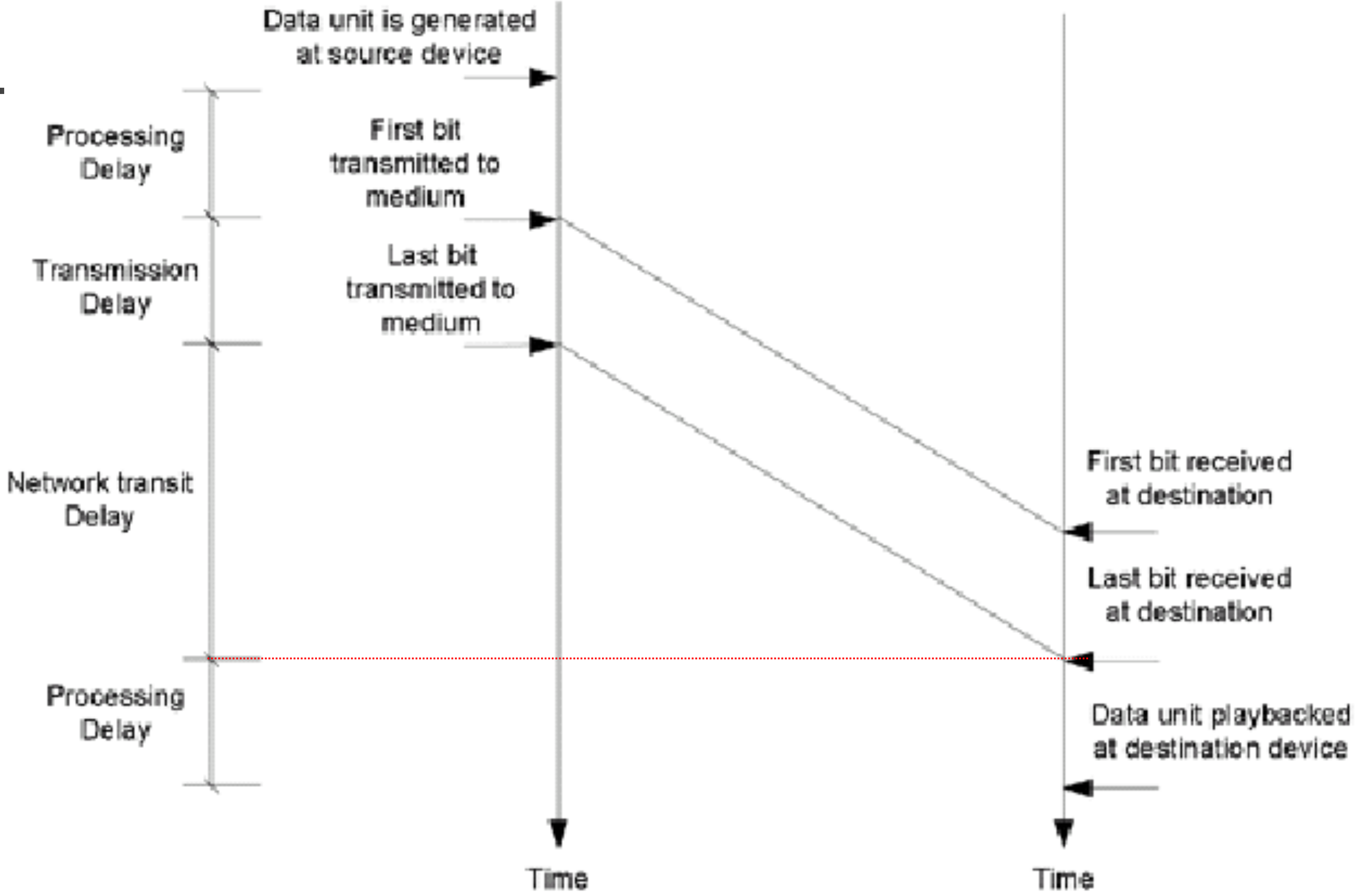
- *burstiness*
  - A measure the degree of bit rate variability of a VBR application.
  - Burstiness=Mean Bit Rate (MBR)/Peak Bit Rate (PBR)
    - PBR: the maximum number of bits in a short period of time
    - MBR the average number of bits in a long period of time
  - *i.e., another definition is Burstiness=PBR/MBR*

Delay

- Delay has a direct impact on users' satisfaction.
- Real-time applications require the delivery of information from the source to the destination within a certain period of time.
- Long delays may cause
  - data missing → reduced video fidelity
  - user frustration during interactive tasks

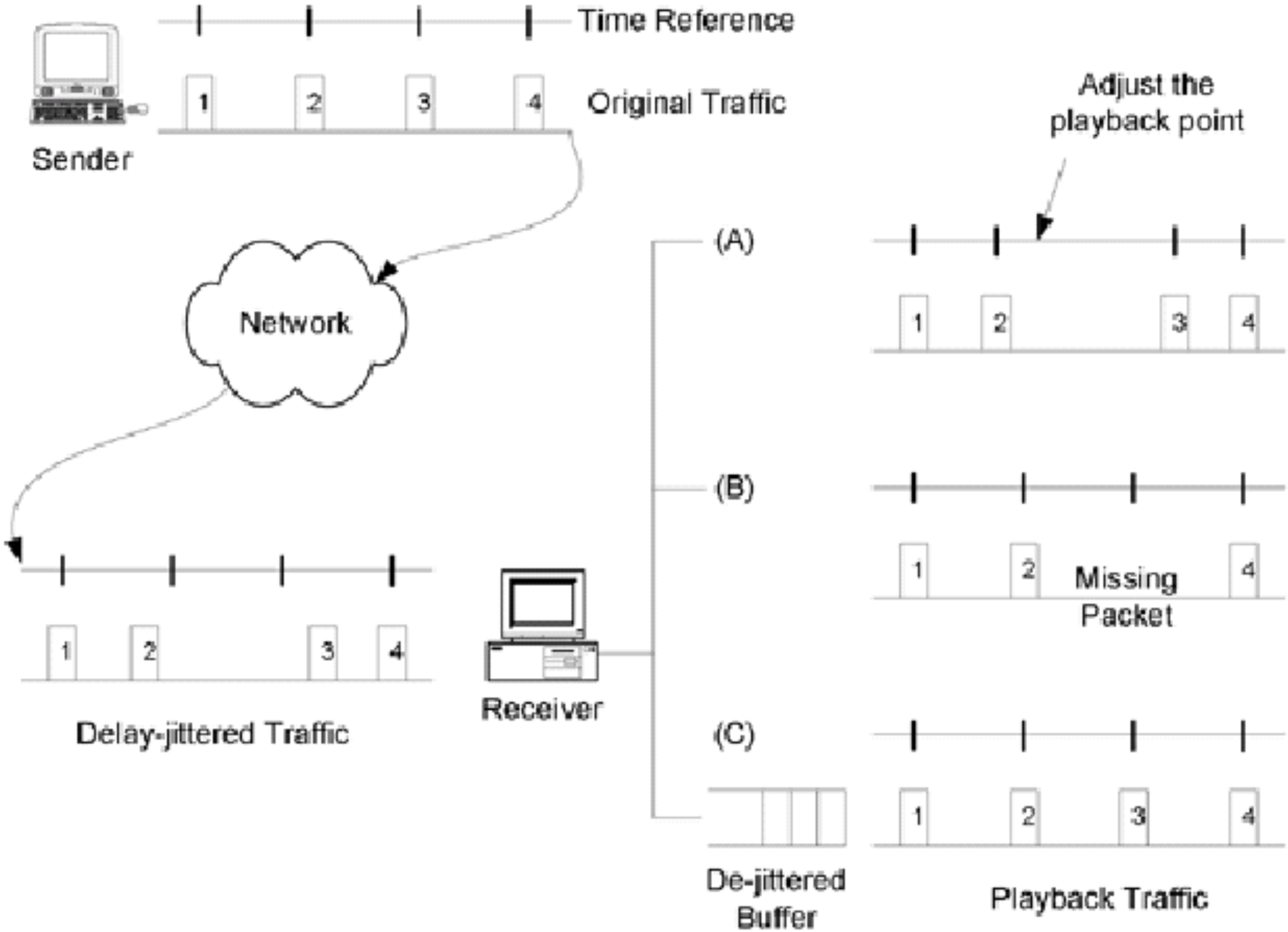
Delay

- Main sources of delay
  - *Source-processing delay* (digitization and packetizing delay):
    - depends on host H/W configuration (CPU power, RAM, motherboard, etc.) and current load (e.g., # of applications running simultaneously).
  - *Transmission delay:*
    - depends on packet size and transmission speed.
  - *Network delay:*
    - *Propagation delay:*
      - depends on the distance between source and destination.
    - *Protocol delay:*
      - depends on protocols, network load, and H/W that executes the protocol.
    - *Output queuing delay:*
      - depends on the network congestion, the configuration of the hardware, and the link speed.
  - *Destination processing delay:*
    - depends on host hardware configuration and load.


Data unit is generated at source device
First bit transmitted to medium
Last bit transmitted to medium
Processing Delay
Transmission Delay
Network transit Delay
Processing Delay
First bit received at destination
Last bit received at destination
Data unit playbacked at destination device
Time
Time

Delay Variation

- Delay variation
  - the variation in the delay introduced by the components along the communication path.


Time Reference
Original Traffic
Sender
1
2
3
4
Adjust the playback point
Network
(A)
(B)
(C)
Missing Packet
Delay-jittered Traffic
Receiver
De-jittered Buffer
Playback Traffic

Delay Variation

- Techniques to cope with delay jitter at the receiver end
  - Technique A:
    - playbacks the signal as soon as the packets arrive (i.e. playback point is changed)
    - Introduces distortion in the playback signal
  - Technique B
    - playbacks the signal based on the original timing reference
    - Introduces distortion (i.e., late packets are ignored)
  - Technique C
    - a de-jittered buffer is used, all packets are stored in the buffer and held for some time (offset delay)
    - Large delay jitter
      - requires large buffer space to hold the packets and smooth out the jitter
      - also introduces large delays, which will be eventually constrained by the application delay requirement

Loss or Error Rate

- Packet loss
  - directly affects the perceived quality of the application.
- Network level, packet loss can be caused by
  - Network congestion
  - Bit errors that occur due to a noisy communication channel.
- Techniques to recover from packet loss or error
  - Error correction at the physical layer
  - Packet retransmission
  - Codec at the application layer

Multimedia Application Requirements

- In many cases, users can determine the application's QoS requirements by investigating the factors that influence the application quality
  - The one-way delay requirements of interactive voice should be less than 250 ms.
- We expect delay and bandwidth bottlenecks to occur in the wireless network due to their limited available bandwidth.

Multimedia Application Requirements

- Factors that influence the application requirements:
  - *Application interactivity level:*
    - Interactive and noninteractive applications
  - *User/Application characteristics:*
    - Delay tolerance and intolerance,
    - Adaptive and nonadaptive characteristics
  - *Application criticality:*
    - Mission-critical and non-mission-critical applications

Interactive & Noninteractive Applications

- Two QoS factors:
  - Delay: the elapsed time between interactions
  - Delay jitter
    - In general, an application with strict delay requirements also has strict delay jitter requirements.
- The degree of interactivity determines the level or stringency of the delay requirement.
  - higher delay tolerance leads to more relaxed delay requirements
  - Ex.
    - interactive voice applications (conversation): ~ ms
    - streaming (playback) video: ~sec.

Tolerance and Intolerance

- Tolerance and intolerance describe the users' sensitivity to changes in QoS parameter values.
- Users' tolerance to latency and distortion:
  - *Latency tolerance and intolerance:*
    - depends on users satisfaction, users expectation, or the urgency of the application
  - *Distortion tolerance and intolerance:*
    - depends on users' satisfaction, users' expectation, and the application media types
      - Ex.: users are more tolerant to video distortion than to audio distortion.

Adaptive & Nonadaptive Characteristics

- Adaptive and nonadaptive aspects mostly describe the mechanisms invoked by the applications to adapt to QoS degradation.

i.e. adaptation is trigged by implicit/explicit feedback from network/end user

Adaptive & Nonadaptive Characteristics

- Common adaptive techniques
  - *Rate adaptation:*
    - *application can* adjust the data rate injected into network
      - data rate can be reduced by
        - dropping some packets,
        - increasing the codec data compression,
        - changing the multimedia properties (i.e., reducing the resolution or color depth of the video).
      - it causes degradation of the perceived quality
  - *Delay adaptation:*
    - *applications* tolerate a certain level of delay jitter by deploying the de-jittered buffer or adaptive playback technique

Application Criticality

- Mission-critical and non-mission-critical aspects reflect the importance of application usage, which determines the strictness of the QoS requirements.
- Failing the mission may result in disastrous consequences.
  - Example:
    - *Remote surgery:*
    - *Telemedicine:* Distorted images may lead to wrong diagnosis.

Representation of Application Requirements

- Two common ways:
  - quantitative expression
    - quantifiable values are determined from
      - application's technical spec. (i.e., video codec—MPEG1, MPEG2, MPEG4, HDTV)
      - Experimentation:
        - *Throughput*: It is mostly expressed as the average data rate:
        - *Delay and delay jitter:* The value is provided in the form of a bound
        - *Loss:* It is mostly expressed as a statistical value
  - qualitative expression
    - Ex. "…. get the bandwidth as much as you can give… " or " … expect low latency…" or " …. provide lower latency than certain applications…."
    - Since these applications do not specify quantitative QoS requirements, it is intractable for networks to provide quantitative services.

Examples of Application Requirements

Table 2.1. Speech Coding Standards

Codec	Bandwidth (kbps)	Sound Quality (MOS)	Codec Complexity
ITU-T G.711	64	> 4	—
ITU-T G.722	48-64	3.8	low
ITU-T G.723.1	6.4/5.3	3.9	high
ITU-T G.726	32	3.8	low
ITU-T G.728	16	3.6	low
ITU-T G.729	8	3.9	medium
GSM 06.10	13	3.5	low
GSM 06.20	5.6	3.5	high
GSM 06.60	12.2	> 4	high
GSM 06.70	4.8 – 12.2	> 4	high

Examples of Application Requirements

Table 2.2. Delay Guidelines for VoIP

One-Way Delay	Effect on Perceived Quality
< 100 – 150 ms	Excellent quality (undetectable delays)
150 – 250 ms	Acceptable quality (slight delays)
250 – 300 ms	Unacceptable quality

Table 2.3. Delay Jitter Guidelines for VoIP

Delay Jitter	Effect on Perceived Quality
< 40 ms	Excellent quality—Undetectable jitter
40—75 ms	Acceptable quality
Over 75 ms	Unacceptable quality

Examples of Application Requirements

Table 2.4. Video Codec Bandwidth Requirements

Video Codec	Bandwidth
Uncompressed HDTV	1.5 Gbps
HDTV	360 Mbps
Standard definition TV (SDTV)	270 Mbps
Compressed MPEG2 4:4:4	25 – 60 Mbps
Broadcast quality HDTV (MPEG 2)	19.4 Mbps
MPEG 2 SDTV	6 Mbps
MPEG 1	1.5 Mbps
MPEG 4	5 kbps – 4 Mbps
H.323 (H.263)	28 kbps – 1 Mbps

QoS Services

- We categorize the QoS services as follows:
  - What kind of service is provided to applications: *quantitative services, qualitative services, best effort services*
  - To which entities (individual or group [class] of applications) the network provides service: *per-flow QoS services, per-class QoS services.*
- Networks may use a combination of QoS services (i.e., per-flow and quantitative, per-class and quantitative).

QoS Services

- Example of QoS services
  - *Guaranteed Integrated Services (IntServ),* which provide per-flow and quantitative QoS service
  - *Controlled Load Integrated Services,* which provide per-flow and qualitative QoS service
  - *Differentiated Services (DiffServ),* which provide per-class and qualitative QoS service

Quantitative (Guaranteed) Services

- Quantitative (guaranteed) services, or services that deliver hard QoS, guarantee the provision of the application quantitative requirements.
  - it delivers the highest quality of service.
  - It guarantees the network performance (i.e., bandwidth, delay, delay jitter) in deterministic or statistical terms.
  - suitable for applications that require quantitative performance guarantee such as mission-critical and interactive applications.
  - a number of QoS mechanisms are required to enable these services.

Qualitative (Differentiated) Services

- Qualitative (differentiated) services, or services that deliver soft QoS, provide relative services.
  - One prominent differentiated services example is the priority service.

Best Effort Services

- Best effort services provide network services without any performance guarantees.
  - all traffic is treated equally.
  - suitable for data traffic (i.e., FTP, email, web pages) that does not require minimum bandwidth or timed delivery.
  - Lowest service level that the network can provide

Per-Flow QoS Services

- Per-flow QoS services provide service assurance to individual flows (applications) quantitatively or qualitatively.
  - Per-flow classification (QoS mechanism that differentiates the flows) is essential to the implementation of per-flow QoS services.

Per-Class QoS Services

- Applications are categorized into different groups (classes) based on different criteria (i.e., QoS requirement, organization, application types, protocol families).
  - Provides service assurance to individual classes quantitatively or qualitatively.
  - Applications in the same class will experience the same QoS.
  - The per-class classification is essential to the implementation of per-class QoS services.

Realization of QoS Services

- Bandwidth (the main network resource) needs to be distributed to applications to satisfy all QoS requirements.
- Two main approaches for bandwidth planning:
  - *Bandwidth over-provisioning:*
    - upgraded the network infrastructure when the current network bandwidth cannot provide QoS support
  - *Bandwidth management:*
    - manage the bandwidth using QoS mechanisms.
    - Examples of QoS mechanisms: classification, admission control, resource reservation, channel access, packet scheduling, and policing.

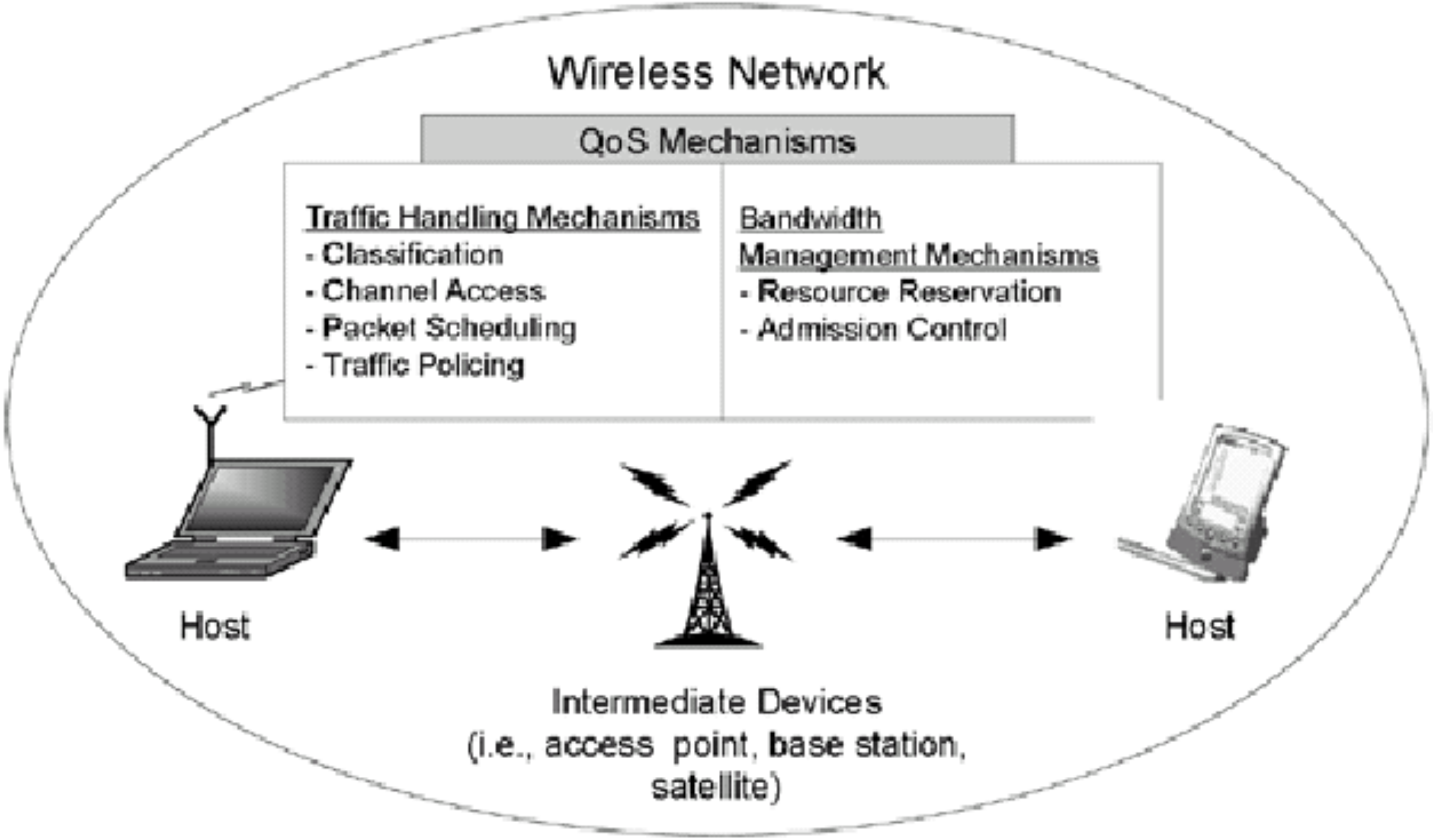
QoS Mechanisms

Outline

- Introduction
- Classification
- Channel Access Mechanism
- Packet Scheduling Mechanisms
- Traffic Policing Mechanism
- Resource Reservation Signaling Mechanisms
- Admission Control
- QoS Architecture

Introduction

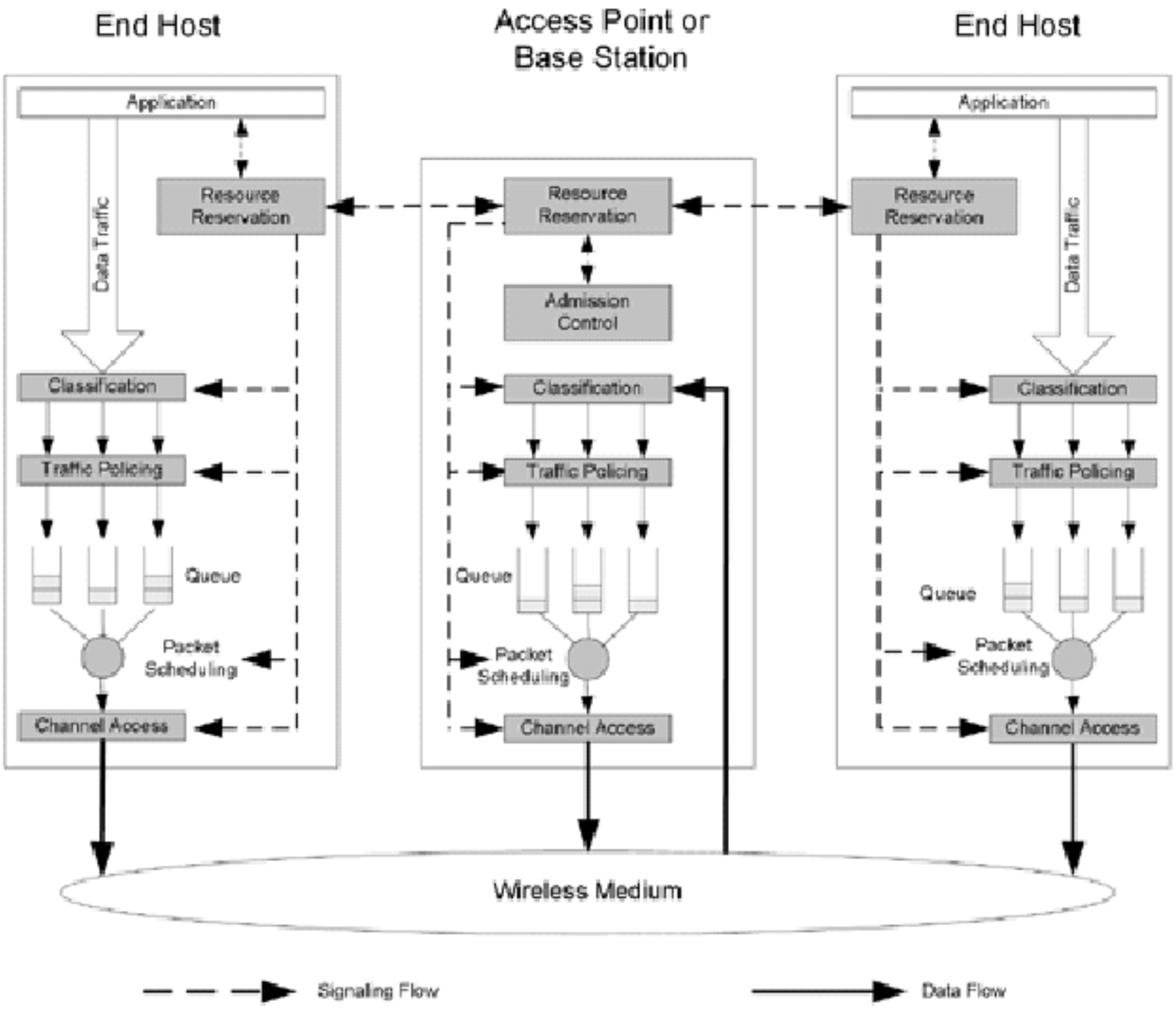
- Two groups of mechanisms:
  - *Traffic handling* (In-traffic) mechanisms
    - classify, handle, police, and monitor the traffic across the network.
    - main mechanisms:
      - classification
      - traffic policing
      - packet scheduling
      - channel access
  - *Bandwidth management* (Out-of-traffic) *mechanisms*
    - mange the network resources (e.g., bandwidth) by coordinating and configuring network devices' traffic handling mechanisms.
    - main mechanisms:
      - resource reservation signaling
      - admission control.


Wireless Network
QoS Mechanisms
Traffic Handling Mechanisms
- Classification
- Channel Access
- Packet Scheduling
- Traffic Policing
Bandwidth Management Mechanisms
- Resource Reservation
- Admission Control
Host
Host
Intermediate Devices
(i.e., access point, base station, satellite)

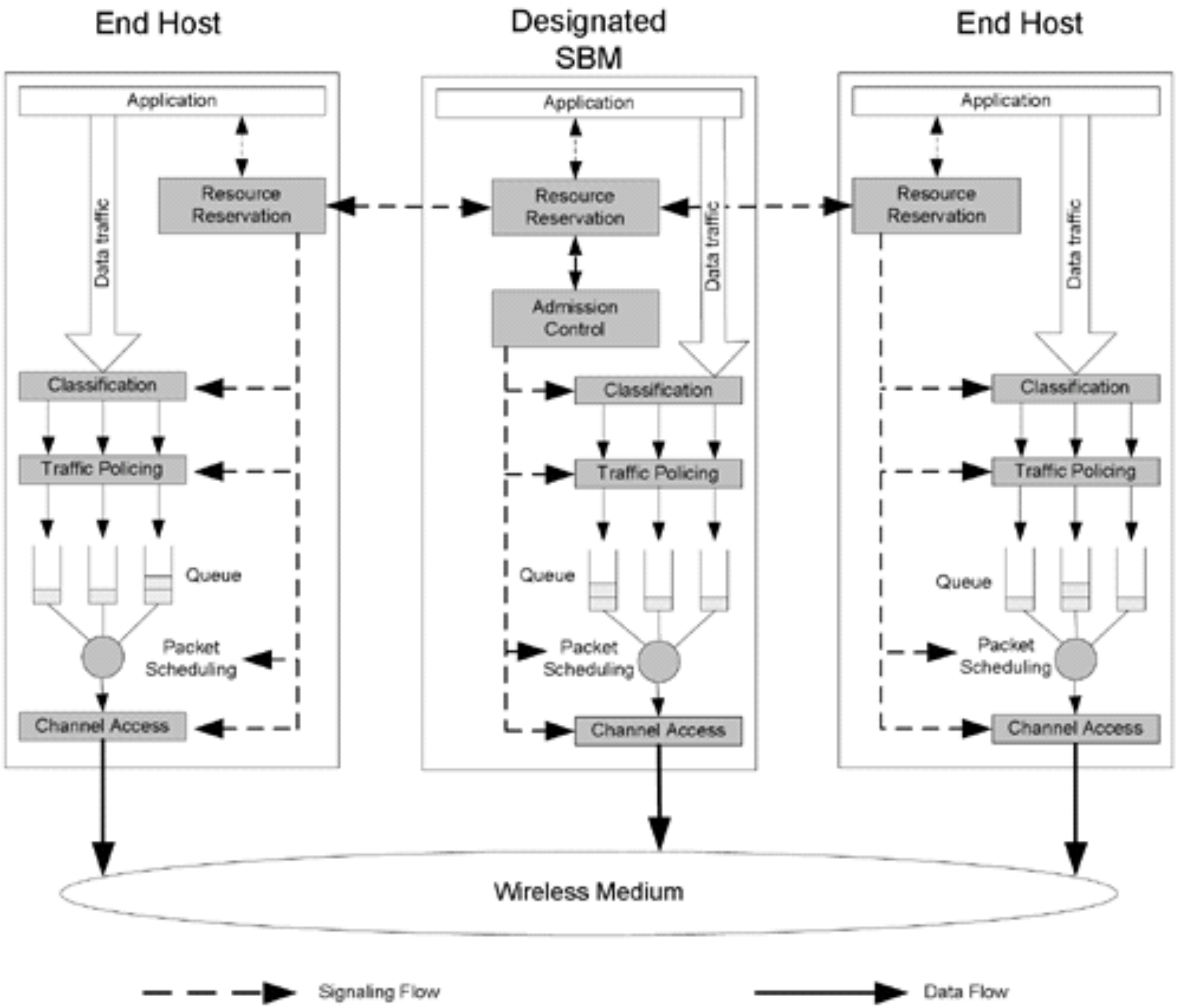
QoS Architecture

- Applications may require different combinations of QoS mechanisms.
  - *Applications with quantitative QoS requirements:*
    - require QoS guaranteed services
    - Requires
      - explicit resource reservation and admission control are needed
      - strict traffic control (traffic policing, packet scheduling, and channel access).
  - *Applications with qualitative QoS requirements:*
    - require high QoS levels but do not provide quantitative QoS requirements.
    - Requires
      - resource reservation and admission control
      - traffic handling which delivers differentiated services
  - *Best effort:*
    - no need for QoS guarantees
    - may reserve bandwidth for best effort traffic

QoS Architecture for Infrastructure Wireless Networks



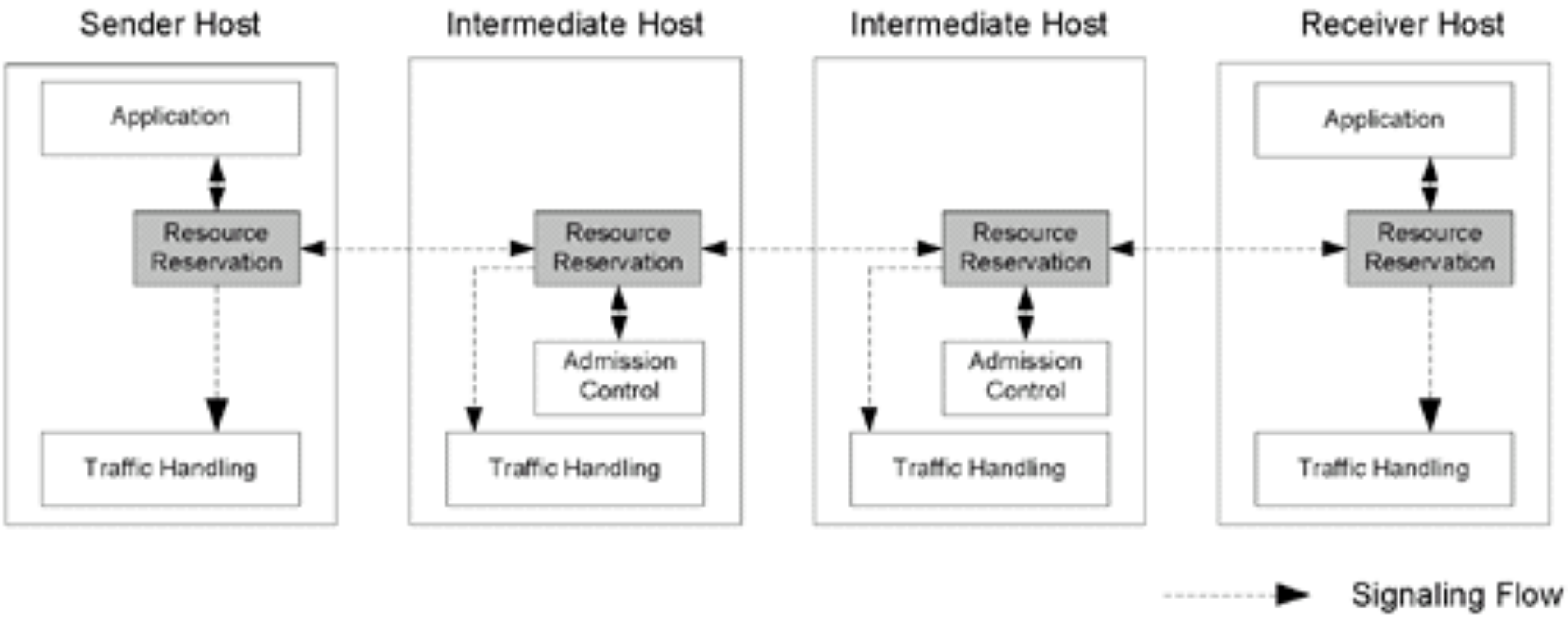
QoS Architecture For Ad Hoc Wireless Networks



Out-of-Traffic Mechanisms

Out-of-Traffic Mechanism

Figure 3.12. Resource Reservation Mechanism



Resource Reservation Signaling Mechanisms

- The traffic handling mechanisms
  - they enable QoS services in each device
  - The mechanisms are classification, channel access, packet scheduling, and traffic policing
- Resource reservation signaling mechanism
  - it inform the network entities on QoS requirements of the multimedia applications
    - The network devices will use this information to manage the network resources (i.e., bandwidth) to meet the requirements.
    - The network devices provide QoS services by configuring the traffic handling mechanisms.
  - it can be applied to individual flows or aggregated flows.
  - it closely cooperates with the admission control mechanisms

Resource Reservation Mechanisms

- Resource reservation mechanisms include following functions:
  - Resource reservation signaling:
    - notifies QoS requirements to all devices along the path
  - Delivery QoS requirements to admission controller:
    - checks available resources
  - Notification of the admission result.
- Resource Reservation Protocol (RSVP)
  - RSVP operates on top of IP, in the transport layer
  - its main functionality is to exchange QoS requirement among devices on the communication path
  - originally designed for supporting per-flow reservation.
  - Currently it is extended to support per-class reservation.

Admission Control

- Admission control
  - a mechanism that decides whether to allow a new session/flow to join the network.
  - ensure that existing sessions' QoS will not be degraded by accepting new session
- Admission control and resource reservation signaling mechanisms are implemented in the same device.

Admission Control

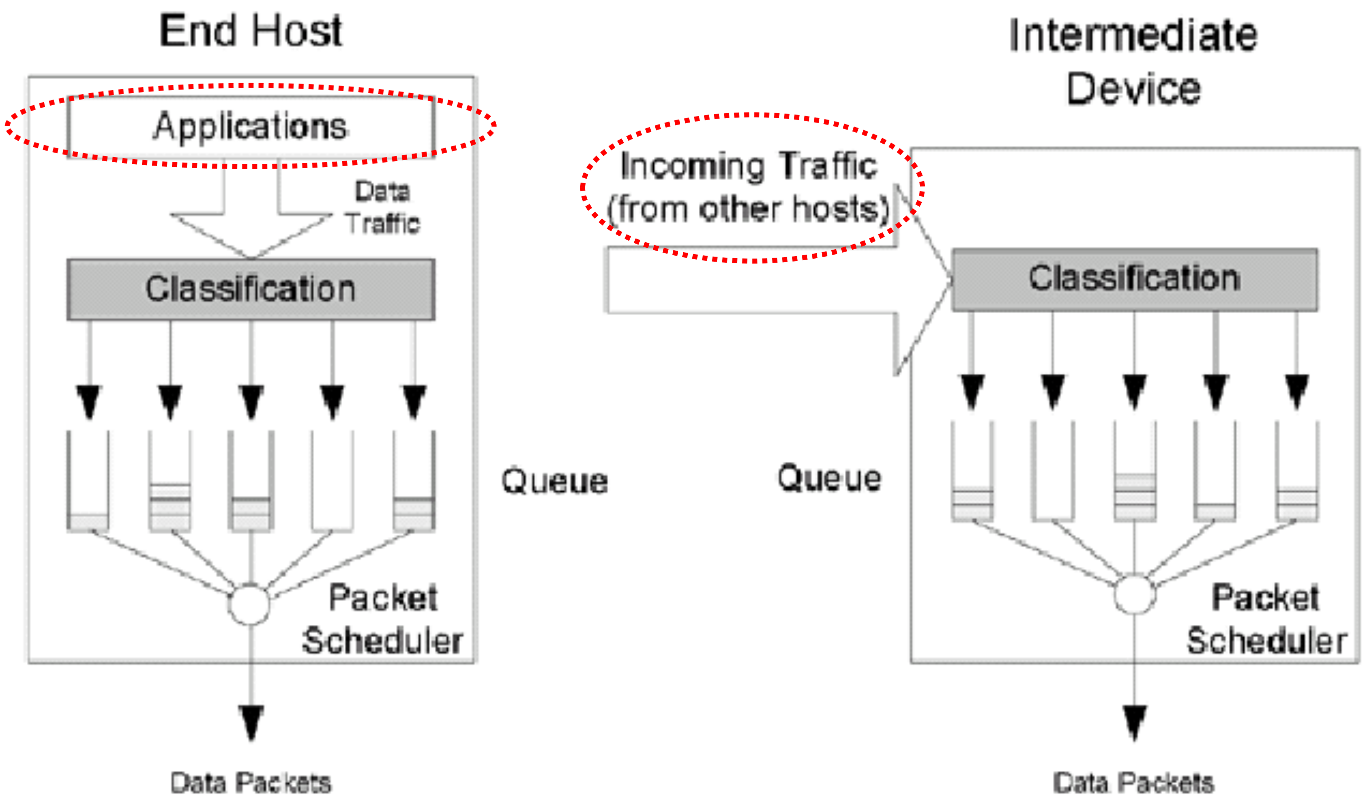
- Two admission control approaches:
  - *Explicit admission control:*
    - based on explicit resource reservation
    - applications will send the request to join the network through the resource reservation signaling mechanism.
    - decision is based on QoS requirements, available resources, performance criteria, and network policy.
  - *Implicit admission control:*
    - no explicit resource reservation signaling
    - the admission control mechanism relies on bandwidth over-provisioning and traffic control (i.e., traffic policing).

In-Traffic Mechanisms

Classification

- Classification:
  - Goal: differentiate between different applications
  - it identifies and separates different traffic into flows or groups of flows (aggregated flows or classes) can be implemented in different network devices
    - it requires some form of tagging or marking of packets to identify and classify the traffic
  - can be per-user, per-flow, or per-class depending on the type of QoS services provided

Classification



Classification

Figure 3.3. Examples of Existing Classification on Each OSI Layer

OSI Layer	Classification Techniques
Application	User/Application Identification
Transport	Flow (5-tuplet IP Address)
Network	IPTOS, DSCP
Data Link	802.1p/Q Classification
Physical Layer	

Data Link Layer Classification

- Data link layer (Layer 2) classifies the traffic based on the tag or field available in Layer 2 header.
  - Example: IEEE 802 user priority
    - IEEE 802 header includes a 3-bit priority field that enables eight priority classes.
    - A classification mechanism identifies packets by examining the priority field and forwards the packets to the appropriated queues.

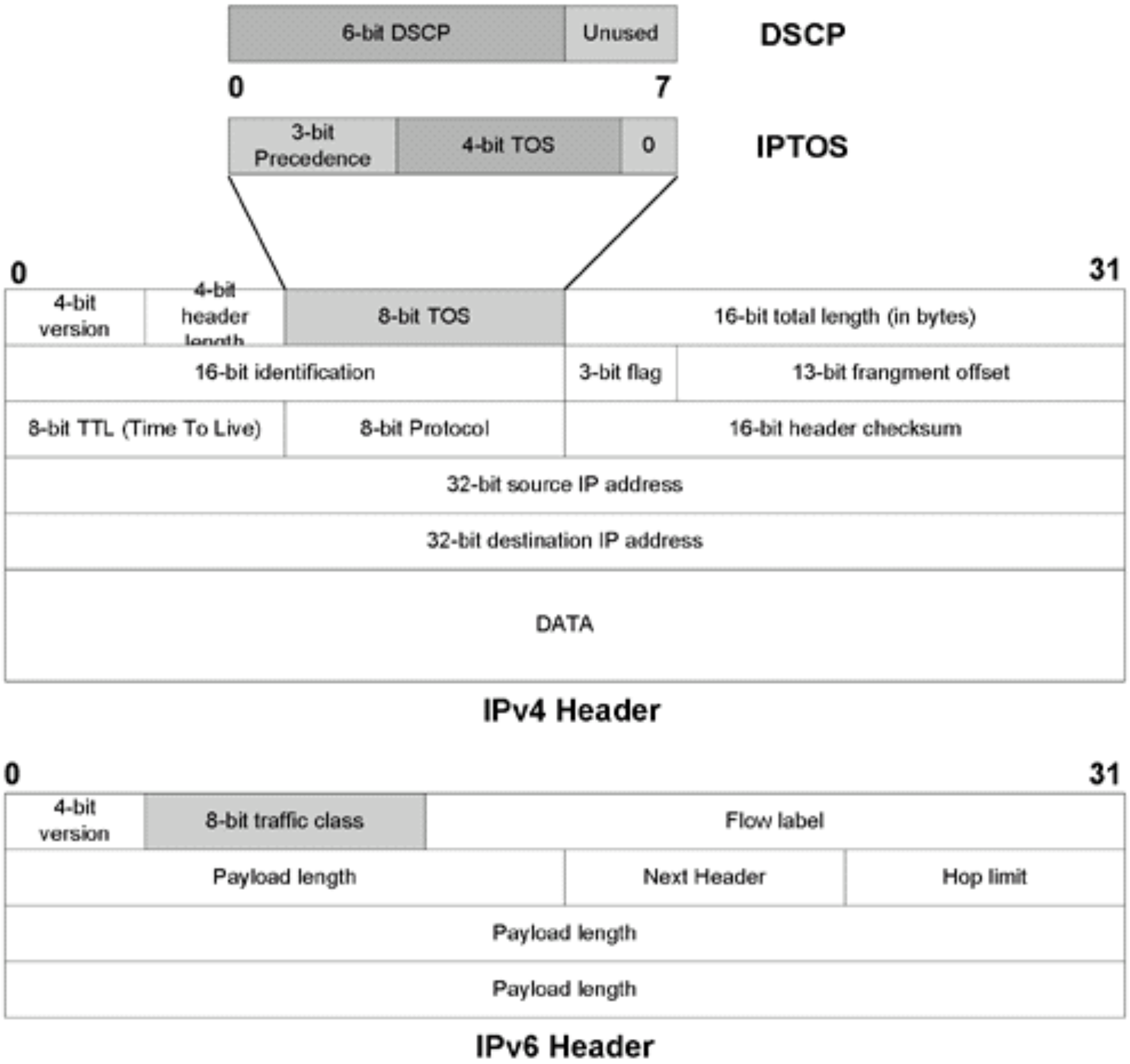
Data Link Layer Classification

Priority	Services
0	Default, assumed to be best effort service
1	Less than best effort service
2	Reserved
3	Reserved
4	Delay sensitive, no bound
5	Delay sensitive, 100ms bound
6	Delay sensitive, 10ms bound
7	Network control

Network Layer Classification

- Network layer (Layer 3) classification classifies packets using Layer 3 header.
  - IPTOS (Internet protocol type of service),
    - IPv4
      - TOS field:
        - a 3-bit precedence subfield,
        - a 4-bit TOS subfield: enables 16 classes of service
        - A final-bit: unused, set to ‘0’
    - IPv6
      - an 8-bit class of service field
  - DSCP (Internet protocol differential service code point).
    - a 6-bit field: enabling 64 classes of service.

Network Layer Classification



Transport Layer Classification

- 5-tuplet IP Header can be used
  - (source IP, destination IP, source port, destination port, and protocol)
  - can uniquely identify the individual application or flow. (i.e., per-flow QoS service can be supported)
  - Limitation
    - Suitable for edge networks, but it is not suitable for core networks because maintaining queues for each individual flow can be an overwhelming task.
    - 5-tuplet IP header exposed to a network outside the firewall that uses NAT (network address translation) cannot uniquely identify the application

Application or User Classification

- The application or user can be uniquely identified by using user/application identification (ID).
- The ID assignment may be static (i.e., the policy or the contract) or dynamic (i.e., connection signaling).
  - Connection signaling,
    - a central station or entity in the network is responsible for making the acceptance decision
    - the application or user sends the connection request
    - the central station assigns a unique ID number if the new connection is admitted
    - Packets from the application will be associated with an ID number.

Traffic Policing Mechanism

- Traffic policing
  - a mechanism that ensures the traffic of admitted sessions do not violate their QoS contract.
    - When violation is found, a traffic policing mechanism is enforced by shaping the traffic.
  - Traffic policing
    - multimedia (real-time) traffic:
      - can provide quantitative traffic parameters
      - traffic policing can be applied to individual flows
    - non-real-time traffic
      - traffic policing enforced (i.e., limits the bandwidth) on aggregated flows based on the network policy.

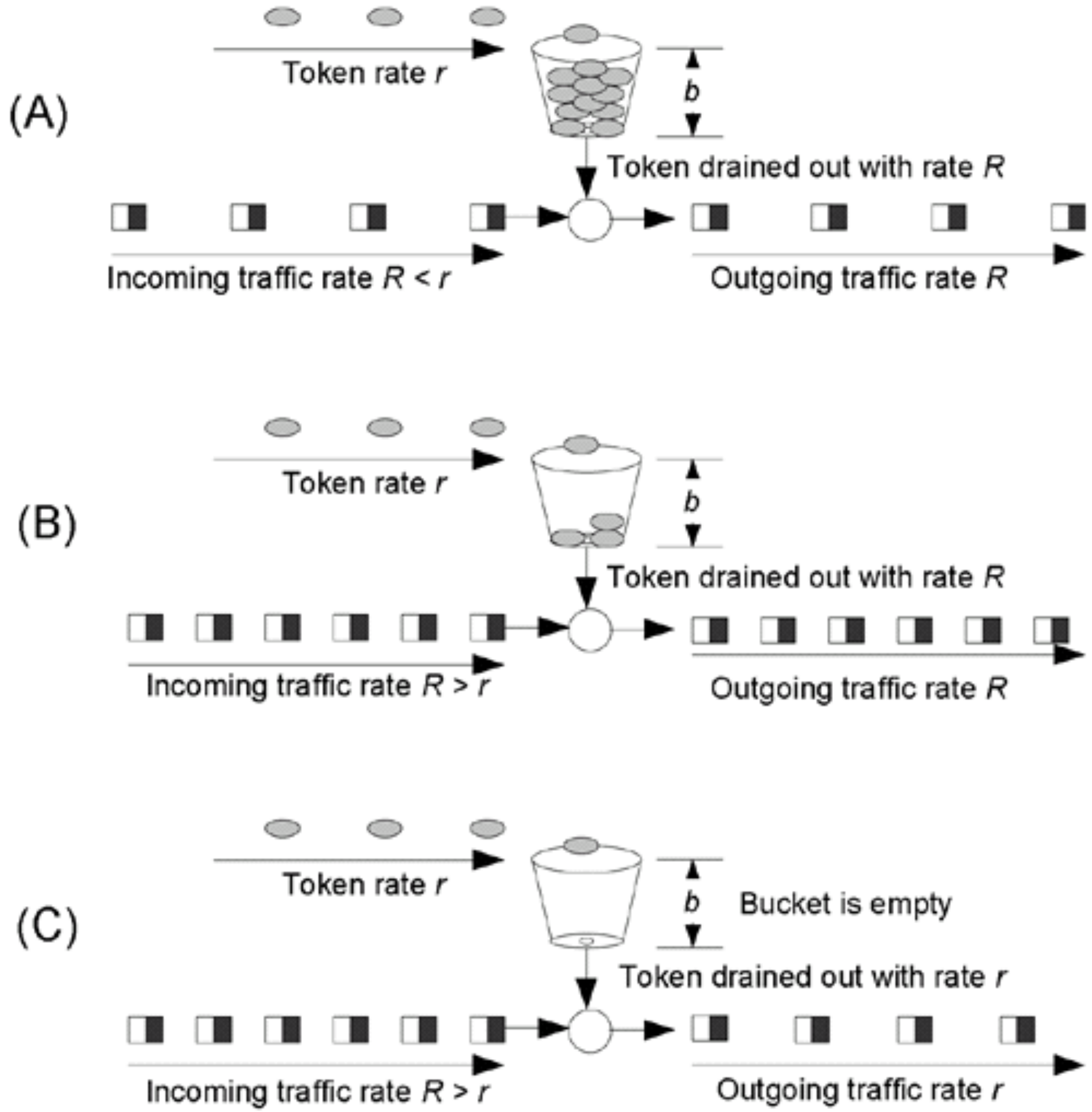
Traffic Policing Mechanism

- Two implementations:
  - Leaky Bucket
  - Token Bucket
- In practical traffic policing,
  - a combination of the token bucket and leaky bucket mechanisms is connected in series (token bucket, then leaky bucket).
    - the token bucket enforces the average data rate to be bound to token bucket rate
    - the leaky bucket (*p*) enforces the peak data rate to be bound to leaky bucket rate.

Token Bucket

- Token Bucket
  - almost the same as the leaky bucket mechanism but it preserves the burstiness of the traffic.
  - Two parameters
    - token bucket of size *b* (bytes)
      - token>0, the outgoing traffic = incoming traffic (rate and pattern)
      - token=0, incoming packets have to wait for tokens
    - tokens filling rate *r* (bytes per second)
  - it preserves the burstiness of the traffic up to the maximum burst size.
  - it is used to control the average rate of the traffic

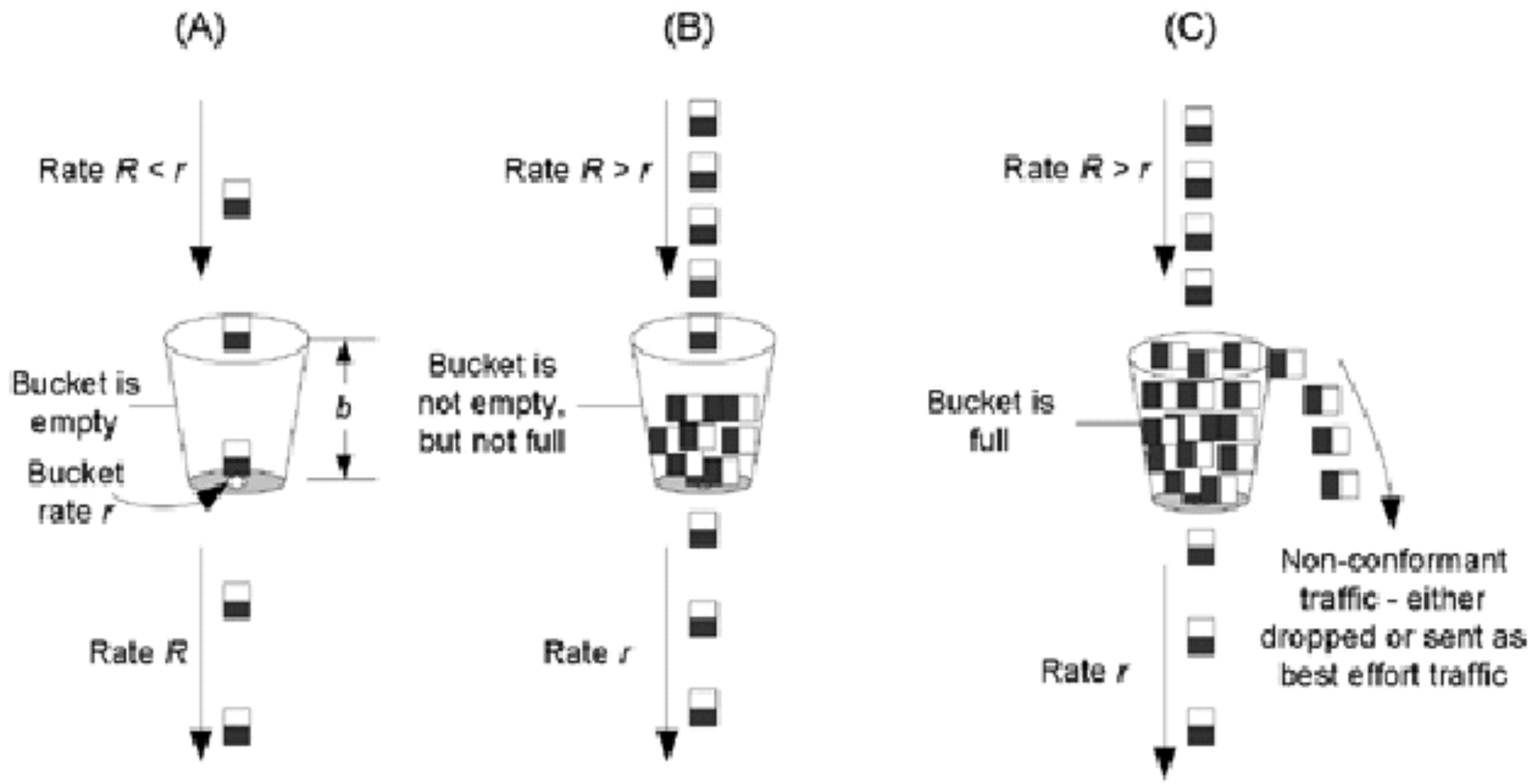
Token Bucket



Leaky Bucket

- Leaky bucket
  - used to smooth the burstiness of the traffic by limiting the traffic peak rate and the maximum burst size.
  - it uses the analogy of a leaky bucket to describe the traffic policing scheme.
  - Two parameters
    - bucket size: maximum burst size
    - hole's size (bucket rate *r*): maximum rate
  - It shapes the traffic with a maximum rate of up to the bucket rate.

Leaky Bucket

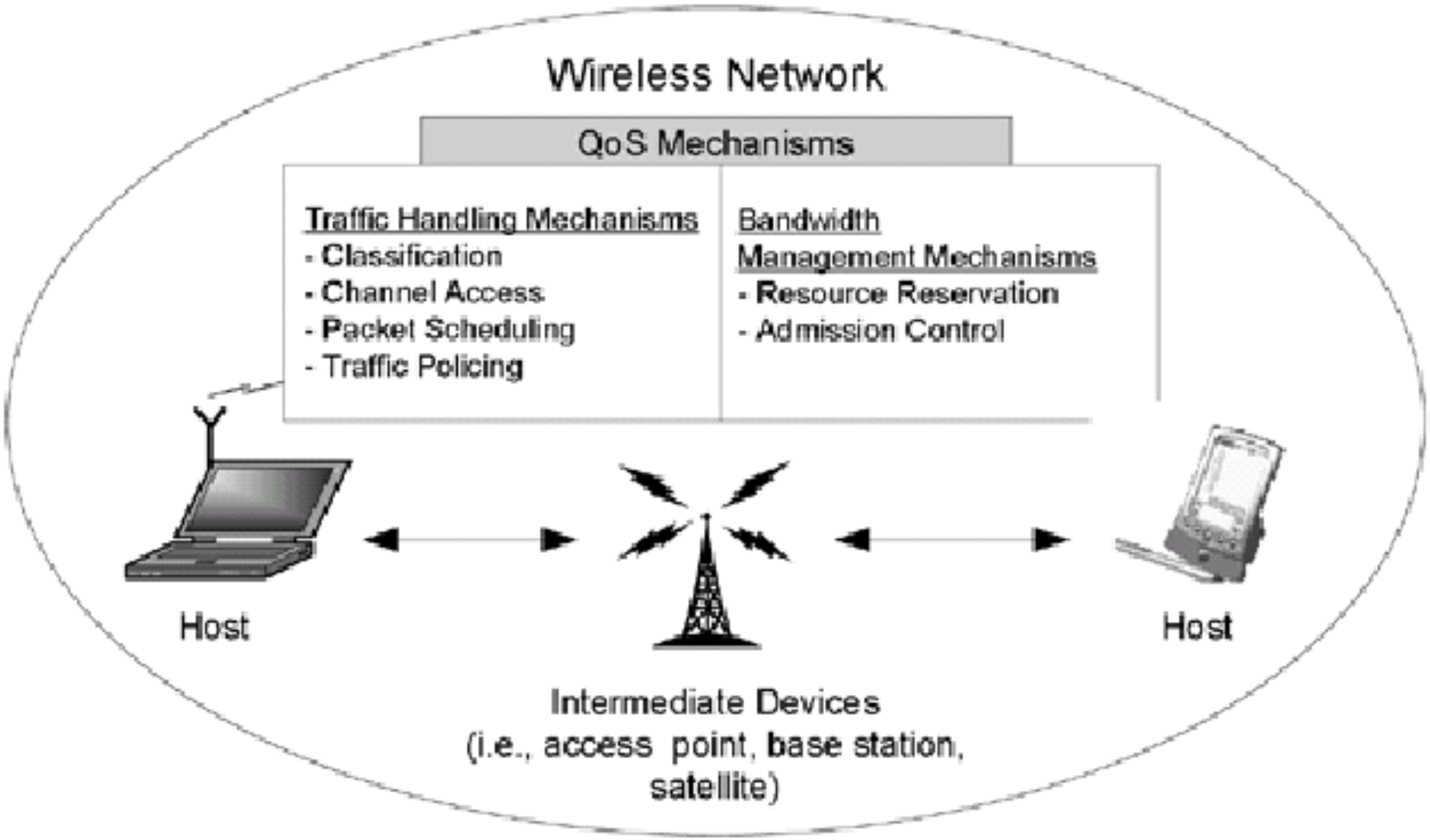


Packet Scheduling Mechanisms

- Packet scheduling
  - it selects a packet from transmission queues for transmission
  - it decides which packet from which queue and station are scheduled for transmission in a certain period of time.
  - it controls bandwidth allocation to stations, classes, and applications
  - It deals with how to retrieve packets from queues, which is quite similar to a queuing mechanism

Packet Scheduling Mechanisms

- Two levels of packet scheduling:
  - *Intrastation packet scheduling:*
    - the scheduling mechanism retrieves a packet from a queue within the same host
    - it is virtually identical to a queuing mechanism.
  - *Interstation packet scheduling:*
    - the scheduling mechanism retrieves a packet from a queue from different hosts.
    - it is slightly different from a queuing mechanism because queues are distributed among hosts.
      - some mechanisms require a signaling procedure to coordinate the scheduling among hosts

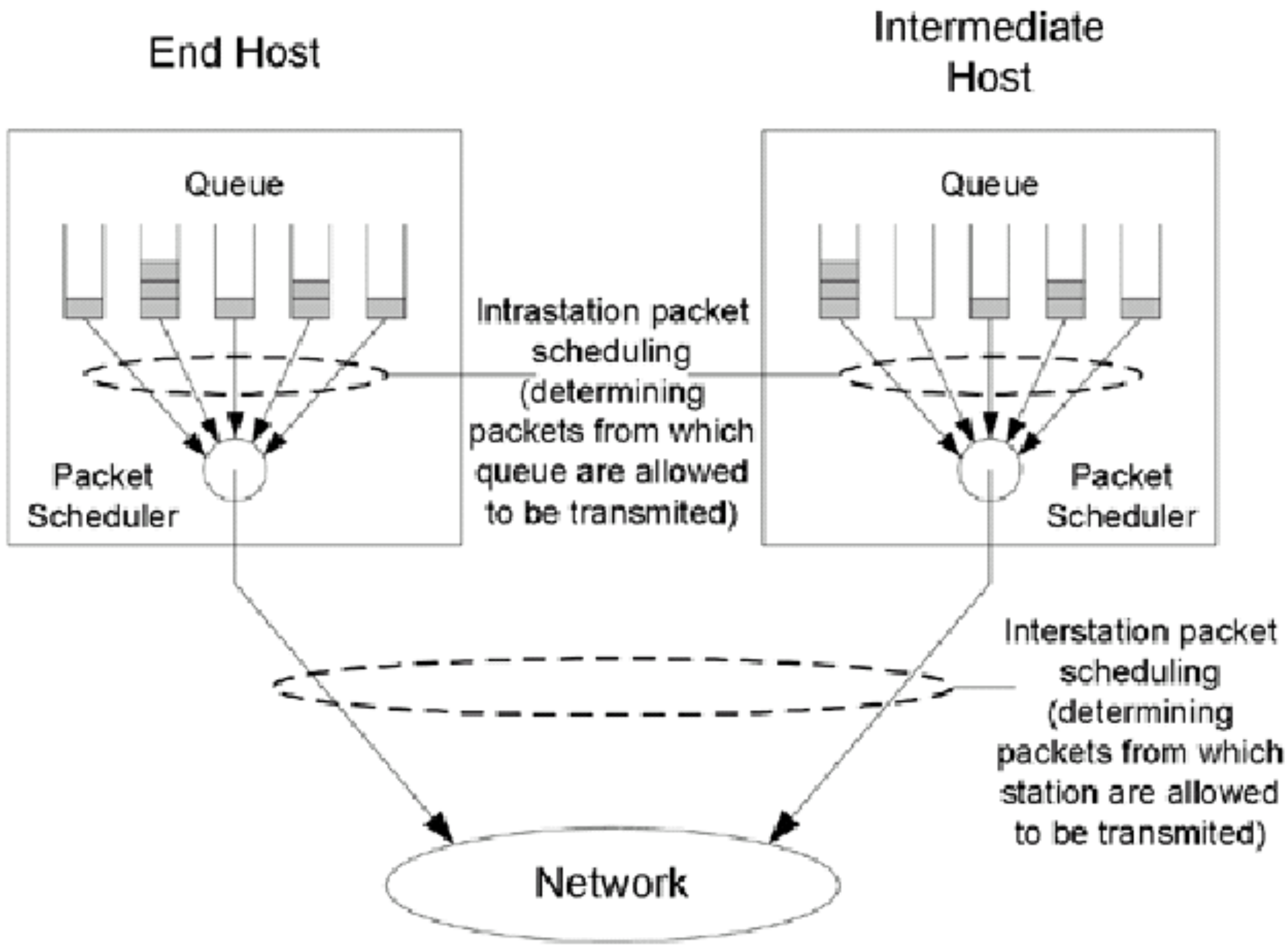

Wireless Network
QoS Mechanisms
Traffic Handling Mechanisms
- Classification
- Channel Access
- Packet Scheduling
- Traffic Policing
Bandwidth
Management Mechanisms
- Resource Reservation
- Admission Control
Host
Host
Intermediate Devices
(i.e., access point, base station,
satellite)

Packet Scheduling Mechanisms

- Two implementation approaches:
  - *Hierarchical packet scheduling:*
    - Bandwidth is allocated to each station
    - When a station receives the opportunity to transmit, the intrastation packet scheduling module will decide which packets to transmit.
    - It is scalable because interstation packet scheduling maintains the state by station (not by connection or application).
    - Overall bandwidth is allocated based on stations (in fact, they Then, stations will have the authority to manage or allocate their own bandwidth portion to applications or classes within the host.
  - *Flat packet scheduling:*
    - Packet scheduling is based on all queues of all stations. Each queue receives individual service from the network.

Packet Scheduling Mechanisms

Figure 3.6. Packet Scheduling



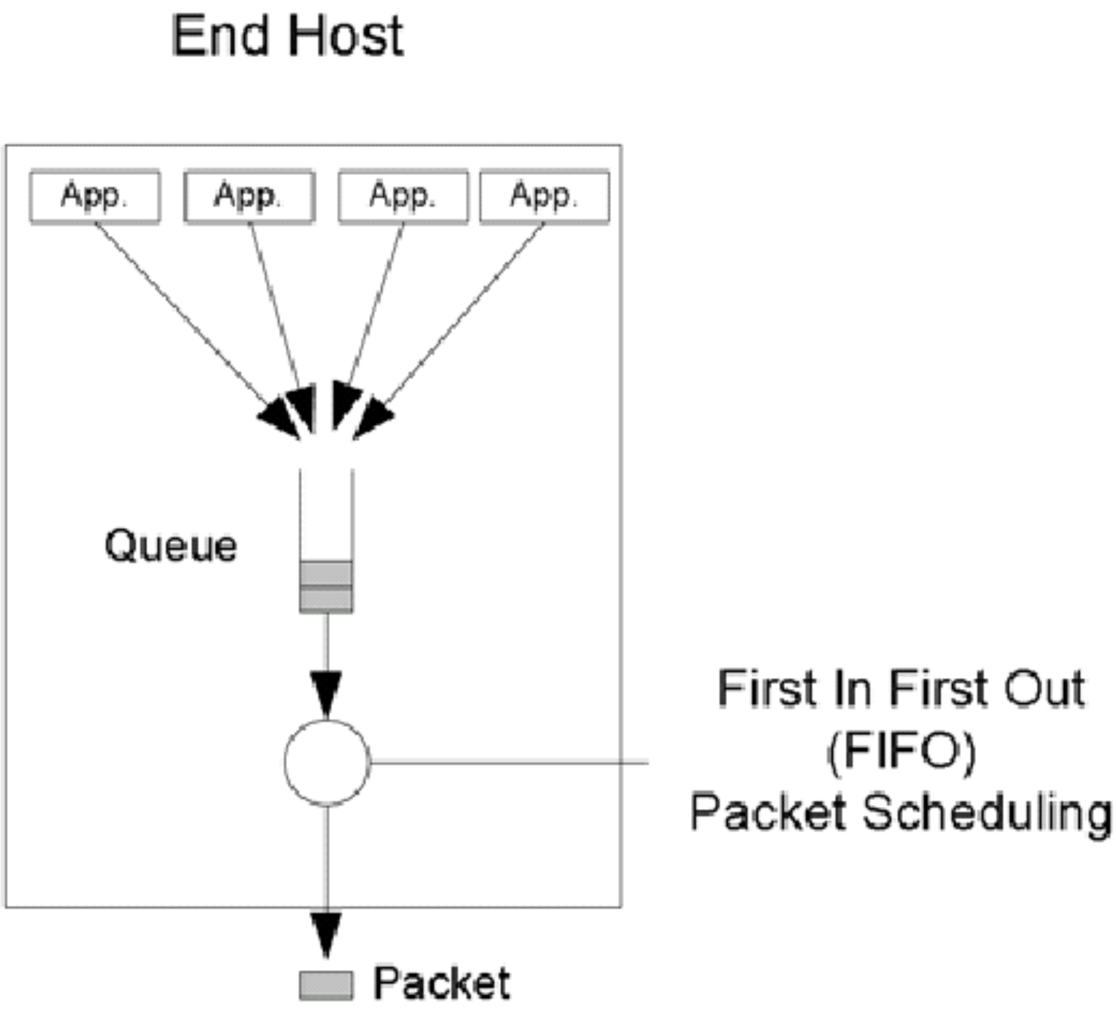
Packet Scheduling Mechanisms

- Queueing schemes
  - First In First Out (FIFO)
  - Strict Priority
  - Weight Fair Queue (WFQ)

Packet Scheduling Mechanisms

- First In First Out (FIFO)
  - Packets are scheduled in order of their arrival.
  - only best effort service can be provided
  - It is possible to improve QoS support by adding
    - traffic policing to limit the rate of each flow
    - admission control

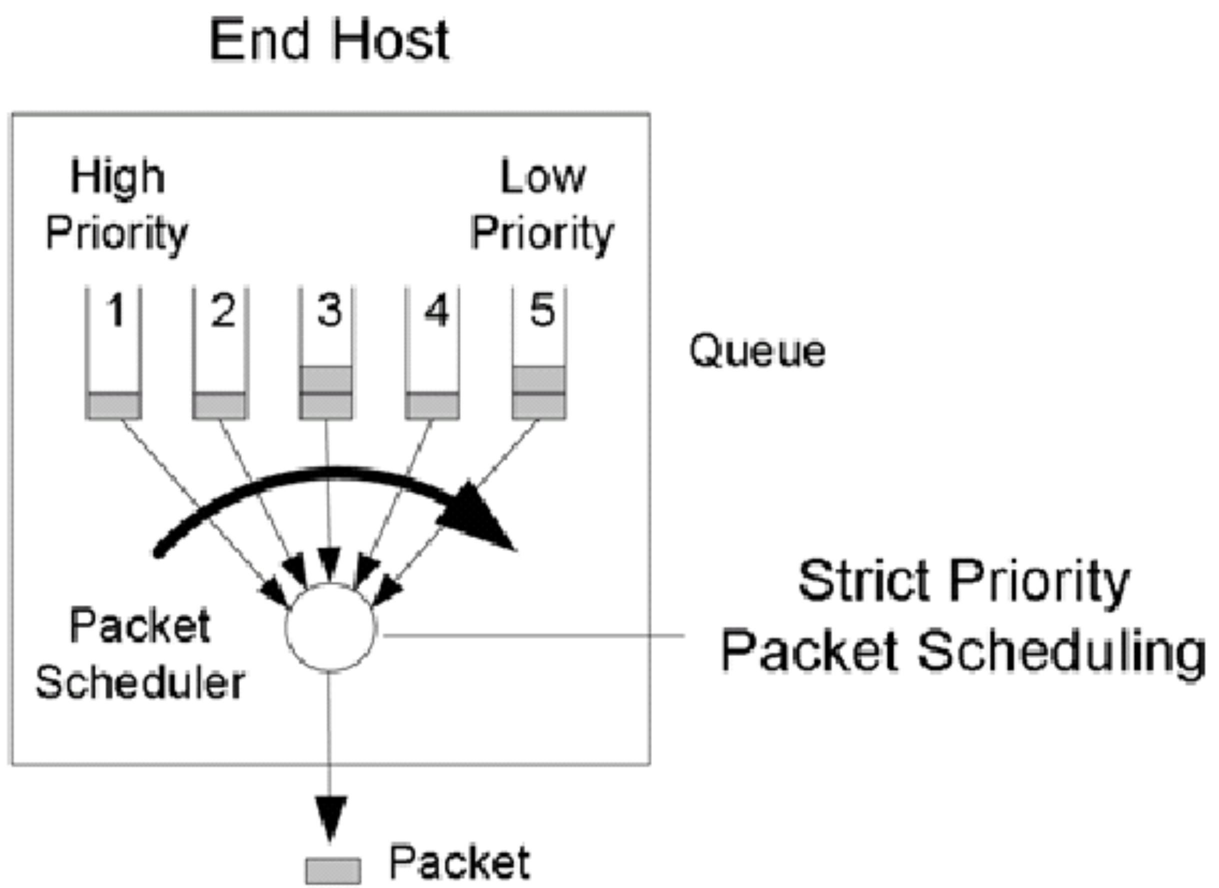
First In First Out (FIFO)



Packet Scheduling Mechanisms

- Strict Priority
  - Strict priority packet scheduling schedules packets based on the assigned priority order.
  - It provides differentiated services (relative services) in both bandwidth and delay.
  - It is possible to improve QoS support by adding
    - traffic policing to limit the rate of each flow
    - admission control

Strict Priority



Packet Scheduling Mechanisms

- Weight Fair Queue (WFQ)
  - it schedules packets based on the weight (wi) of each queue. wi is assigned to queue i according to the network policy.
  - Bandwidth abuse from a specific queue won't affect other queues.
  - WFQ can provide the required bandwidth and the delay performance is directly related to the allocated bandwidth.
  - WFQ mostly schedules packets that belong to aggregated flows, groups, and classes (instead of individual flows) where the goal is to provide link sharing among groups. In this case, delay is of less concern.

Channel Access Mechanism

- In wireless networks, all hosts communicate through a shared wireless medium.
- When multiple hosts try to transmit packets on the shared communication channel, collisions can occur.
- A *channel access mechanism* is required to control the access to the shared channel.
- Two types of channel access mechanisms:
  - collision-based channel access
  - collision-free channel access

Collision-Based Channel Access

- Collision-based channel access
  - a distributed channel access method that provides mechanisms to avoid and resolve collisions
  - collision probability depends on the number of active users
    - Traffic load↑, number of collisions and retransmissions ↑, delay ↑
    - the number of collisions and re-transmissions is random, it leads to an unbound delay
  - collision-based channel access schemes can provide best effort service

Collision-Based Channel Access

- The service level can be improved by:
  - Over-provisioning
    - all traffic will receive ample (豐富的) of bandwidth and experience low delay
  - Adding a priority scheme in the collision-based channel access
    - using different sized backoff windows for different priority classes.
    - it enables the provision of differentiated services.

Collision-Free Channel Access

- Collision-free channel access mechanism
  - channel is arbitrated such that no collision can occur.
  - only one host is allowed to transmit packets to the channel at any given time.
  - Examples
    - polling
    - TDMA (Time Division Multiple Access).
  - provides tight channel access control that can provide a tight delay bound.
  - good candidates for QoS provision to applications with strict QoS requirements

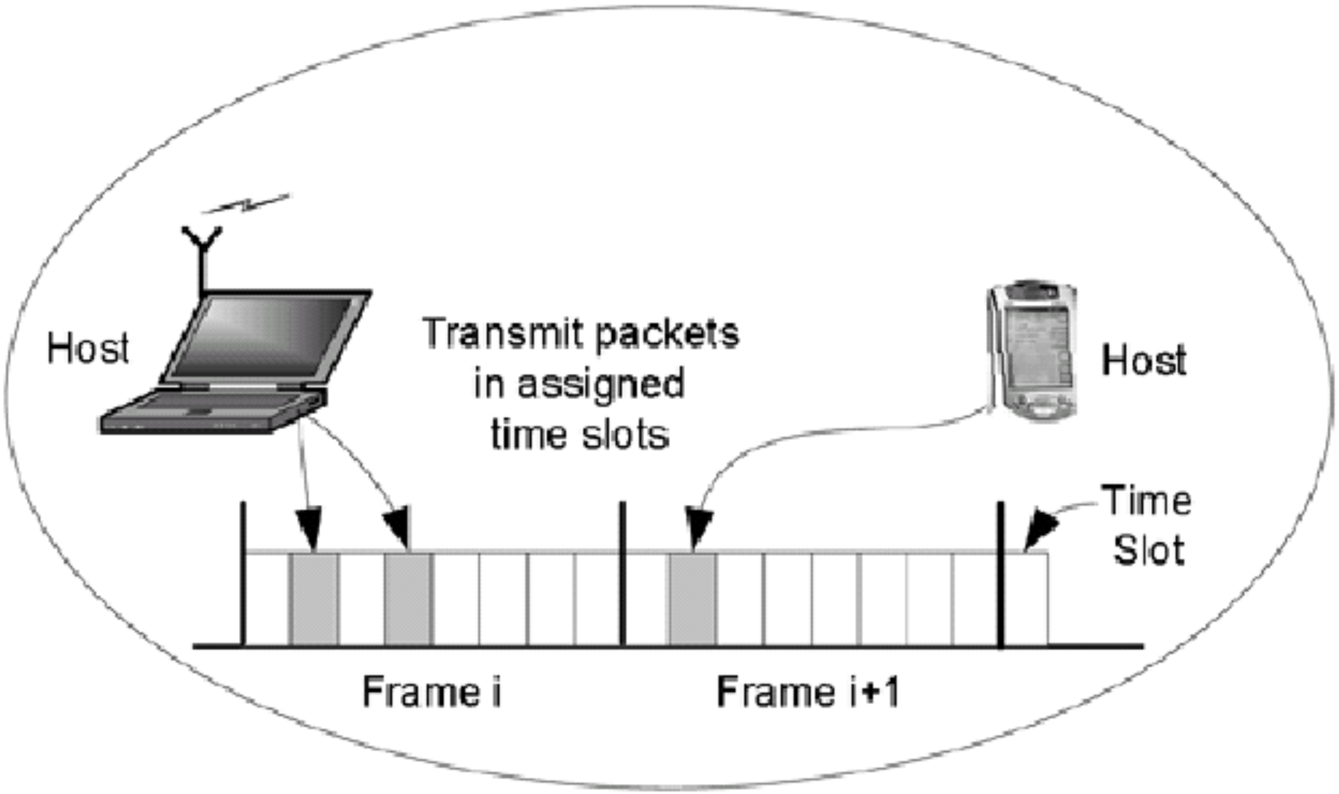
Polling

- A host in the network is designated as the poller, which controls all access to the wireless channel by the other hosts (i.e., pollees).
- Pollees are not allowed to transmit packets unless they receive a polling packet from the poller.
- No collision.
- The polling frequency (the number of polls in a period of time) reflects the bandwidth allocation.
- Bandwidth can be dynamically allocated by adjusting the polling frequency.

TDMA

- A TDMA scheme
  - it divides the channel access opportunity into frames and each frame is divided into time slots
  - a host is allowed to transmit packets in a predefined time slot
  - number of time slots assigned to a host per frame reflects the bandwidth allocated for the host
  - it requires a master host to manage the time slot assignment for all the hosts in the network

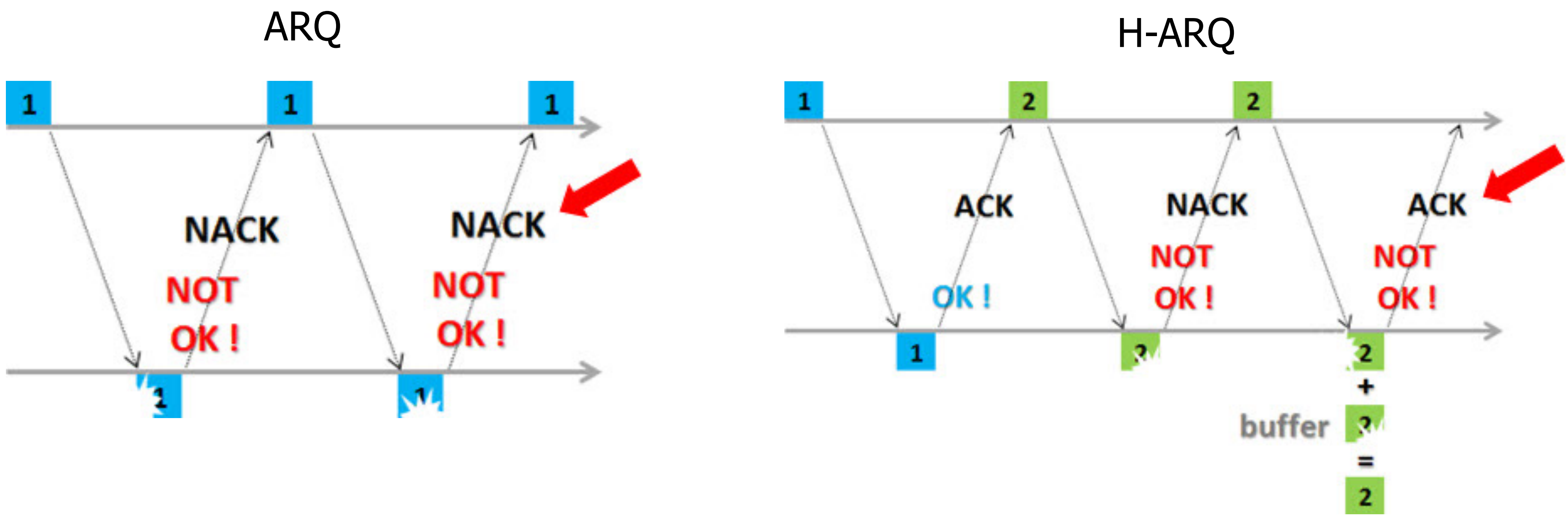
TDMA


Host
Transmit packets in assigned time slots
Host
Time Slot
Frame i
Frame i+1

TDMA

- Time slot assignment philosophies:
  - *Static time slot assignment:*
    - Each host receives a fixed time slot assignment which can be provided during the connection setup.
  - *Dynamic time slot assignment:*
    - time slot assignment changes dynamically during the lifetime of the session
    - Time slot assignment is a function of the traffic load, application QoS requirements, and channel conditions
    - more flexible and leads to better channel utilization
    - extra signaling overhead is required

ARQ v.s. HARQ



ARQ
1
1
1
NACK
NOT
OK !
NACK
NOT
OK !
H-ARQ
1
2
2
ACK
OK !
NACK
NOT
OK !
ACK
NOT
OK !
1
2
+
buffer
=
2

Channel Access

- **channel access/multiple access scheme**
  - allows several terminals connected to a shared medium to transmit over it and to share its capacity.
  - a channel-access scheme is based on multiplexing method, that allows several data streams or signals to share the same communication channel or physical medium.
  - a channel-access scheme is also based on a multiple access protocol and control mechanism, also known as media access control (MAC). This protocol deals with issues such as addressing, assigning multiplex channels to different users, and avoiding collisions.

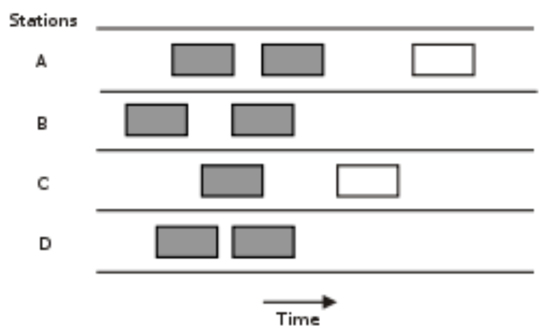
Channel Access

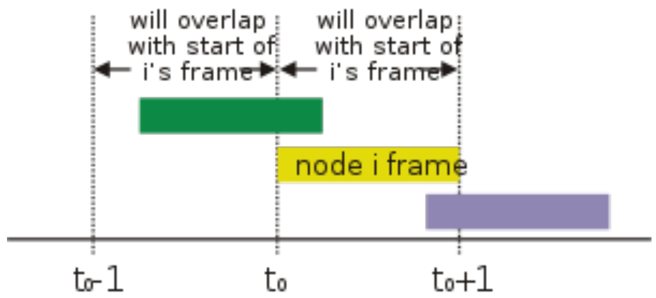
V · T · E	Channel access methods and Media access control (MAC)/Multiple-access protocols [hide]	
Channel based	FDMA	OFDMA · WDMA · SC-FDMA
	TDMA	MF-TDMA · STDMA
	CDMA	W-CDMA · TD-CDMA · TD-SCDMA · DS-CDMA · FH-CDMA · OFHMA · MC-CDMA
	SDMA	HC-SDMA
	PDMA	
	PAMA	
Packet based	Collision recovery	ALOHA · Slotted ALOHA · R-ALOHA · AX.25
	Collision avoidance	MACA · MACAW · **CSMA** · CSMA/CD · CSMA/CA · DCF · PCF · HCF · CSMA/CARP
	Collision free	Token ring · Token bus · MS-ALOHA
	Delay & disruption tolerant	DTN · Mobile Ad-Hoc · Dynamic Source Routing
Duplexing methods	TDD · FDD	

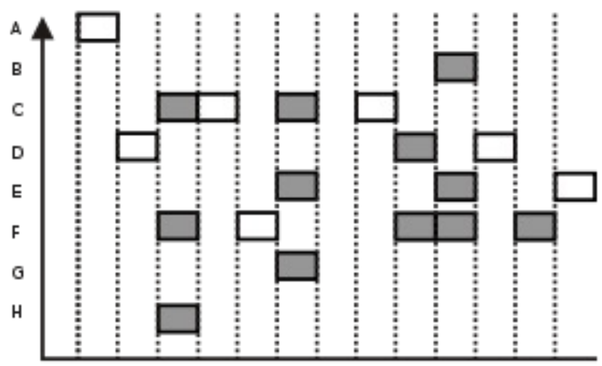
SDMA: Space-division multiple access

Carrier sense multiple access (CSMA)

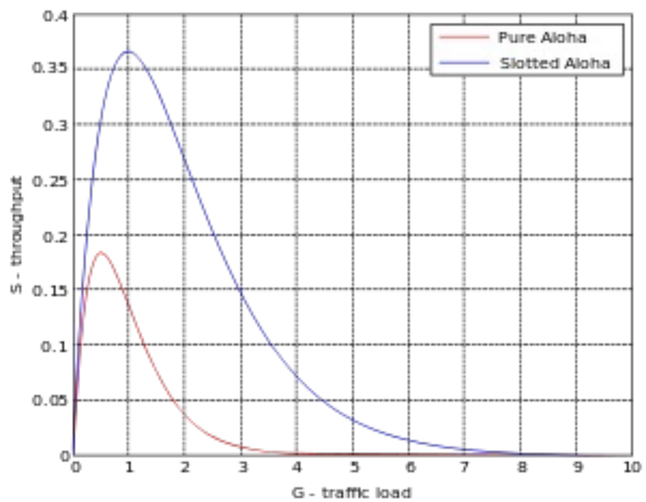
ALOHA/Slotted ALOHA

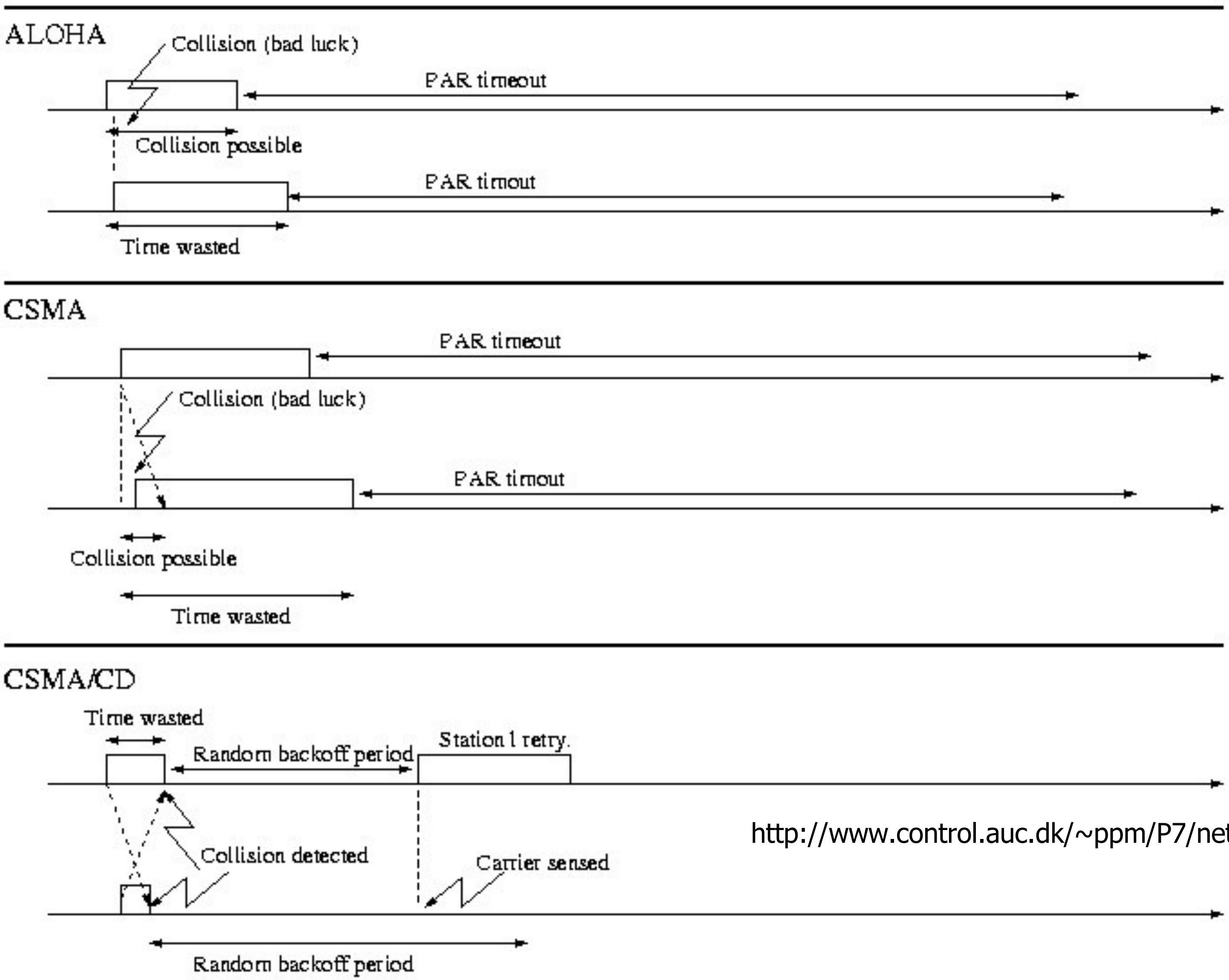






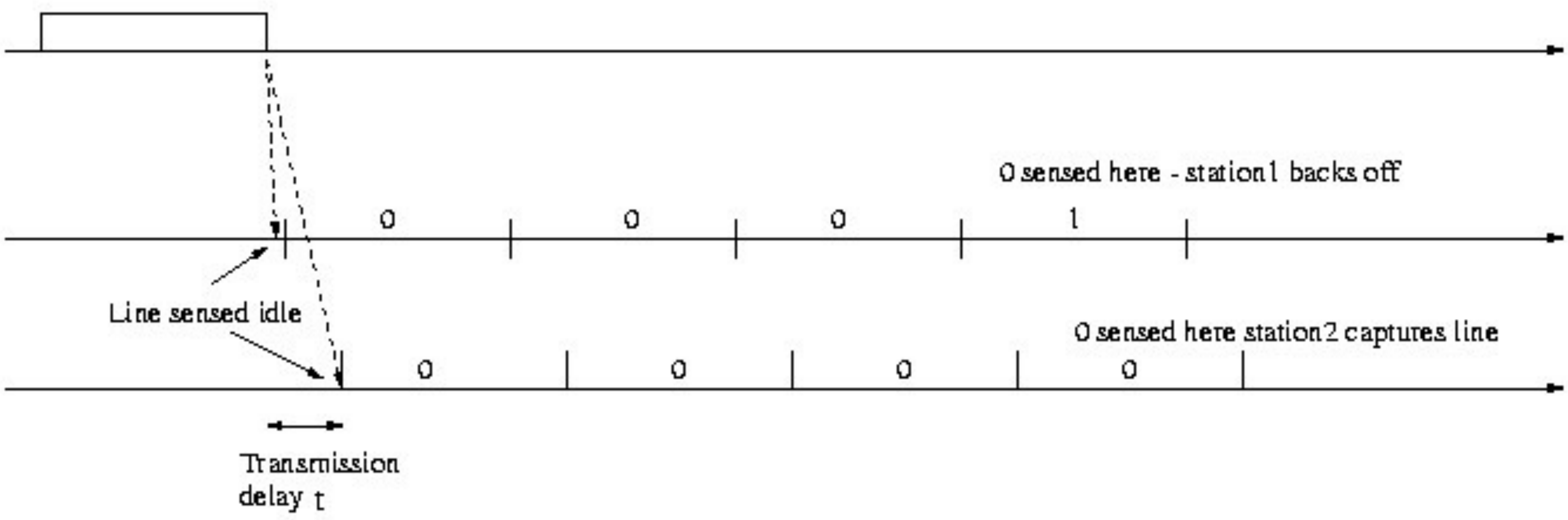
Slotted ALOHA protocol (shaded slots indicate collision)





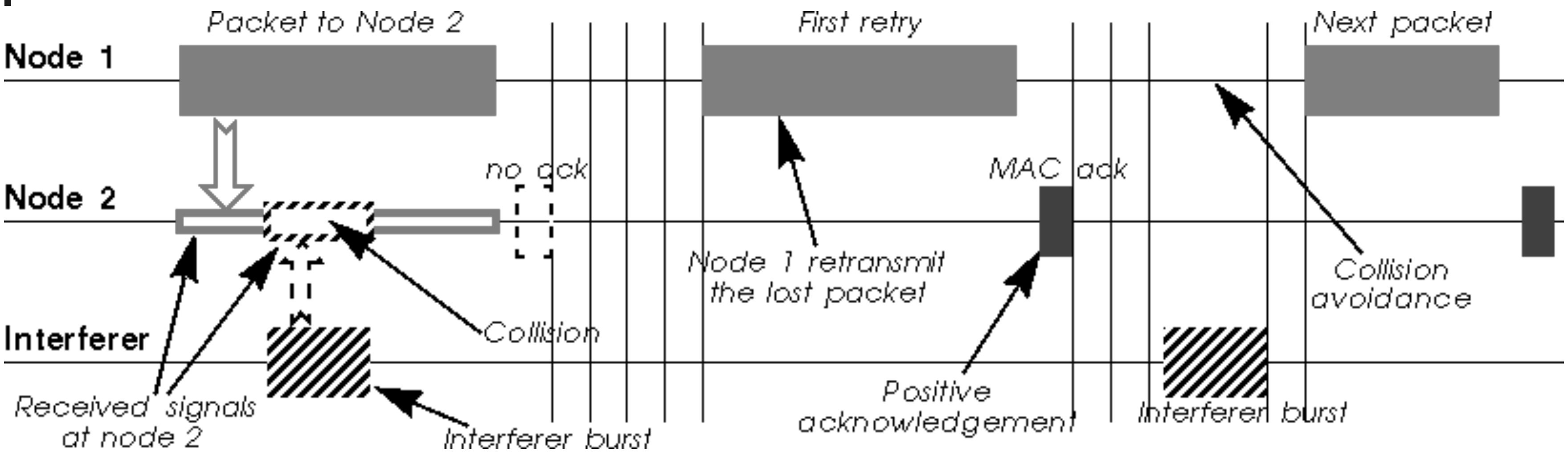
http://www.control.auc.dk/~ppm/P7/netwaerk/note_mm2.html

CSMA/CA (Collision Avoidance)

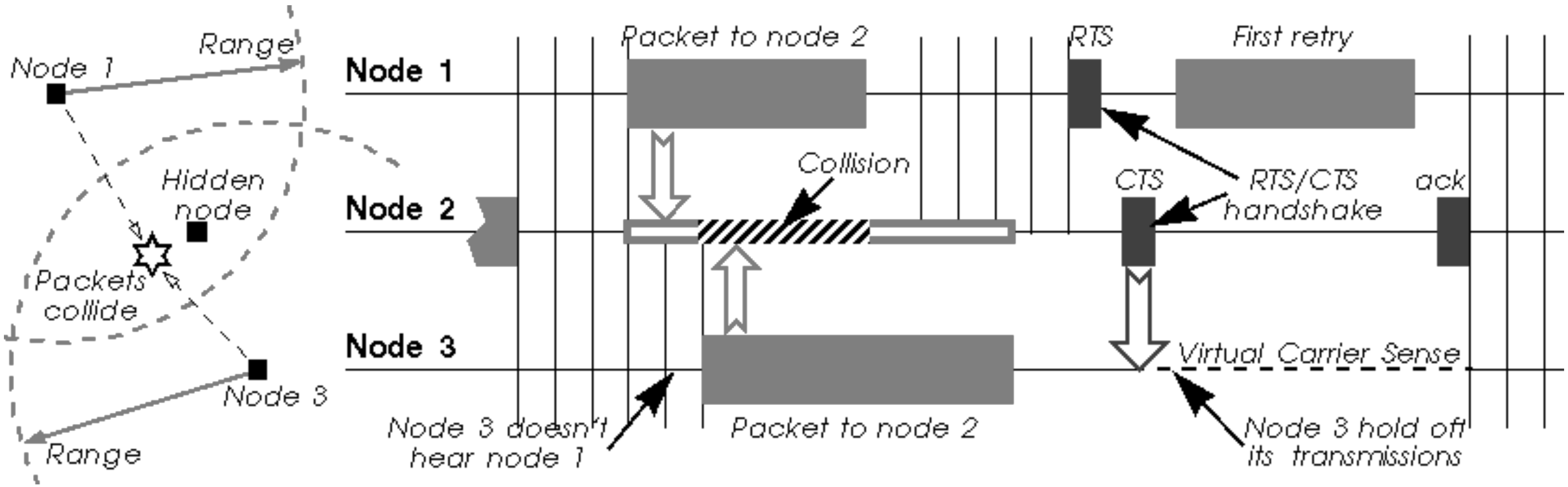


Bitlength >> t

CSMA/CA



Packet to Node 2
First retry
Next packet
Node 1
Node 2
no ack
MAC ack
Interferer
Node 1 retransmit the lost packet
Collision avoidance
Collision
Received signals at node 2
Interferer burst
Positive acknowledgement
Interferer burst


Node 1
Range
Hidden node
Packets collide
Node 3
Range
Node 1
Node 2
Node 3
Packet to node 2
Collision
Node 3 doesn't hear node 1
Packet to node 2
RTS
First retry
CTS
RTS/CTS handshake
ack
Virtual Carrier Sense
Node 3 hold off its transmissions