ICC2001

The IEEE International
Conference on Communications

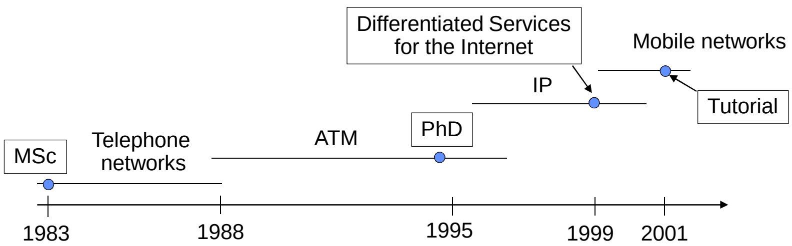
Tutorial 14, June 11, 2001

Service Differentiation in the Mobile Internet

Dr. Kalevi Kilkki
Nokia, Finland

Service Differentiation in the Mobile Internet

- by Kalevi Kilkki (kalevi.kilkki@nokia.com)
  - previously at Helsinki University of Technology, and Sonera
  - now as a Principal Scientist at Nokia Research Center, Helsinki
    - but all presented opinions are personal
  - main interest: quality of service



Objective of the Tutorial

- Key concepts
  - service differentiation ⇔ business of service provider
  - mobile Internet ⇔ mobile technology
- Primary goal is to clarify the relationship between business and technology
- Concentration on essential matters
  - technical details are ignored unless they have a clear effect on the service differentiation
- In the end of day
  - you should understand the fundamental principles of service differentiation, and
    how those principles can be applied to mobile networks

Contents

- In order to compare any systems (e.g., QoS)
  we need a general framework
  - with clearly defined objectives and criteria for comparison
  - here
    - objective = service provider's goal, usually business
    - tool to define the criterion and to make analysis = utility
- Technical system
  - radio access networks in general
  - with notes on specific systems (GPRS, UMTS, WLAN)
  - emphasis on service differentiation over radio access
- A case study
  - ISP Goals & Utility & Radio Network ⇒ ?

Timetable

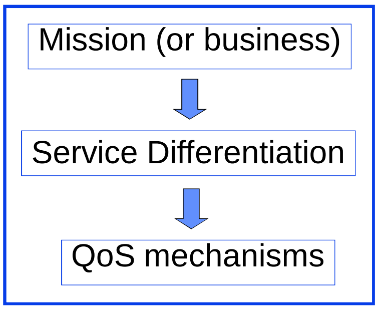
start	duration	topic
13:30	0:10	Introduction
13:40	0:50	Objectives and tools
14:30	0:20	Coffee
14:50	0:20	Radio networks
15:10	0:40	GPRS, UMTS, WLAN
15:50	0:10	Break
16:00	0:40	Case study (imaginary)
16:40	0:20	Conclusion & discussion
17:00		The End

Session 1, Part A

Definition of Objectives

What is service differentiation?

- *Service differentiation is a tool used by service providers to support their mission*
  - e.g., the profitability of their business or the wealth of the society
- Framework



Mission is the starting point and reference for all other matters,

service differentiation shall serve the mission,

QoS mechanisms shall serve the service differentiation

Applications´ role in the framework

- The basic assumption used here is that

> The objective of
> Mobile Internet Service Provider (M-ISP)
> is profitable business
>
> *due to the extremely high license fees*
> *it is very difficult to assume anything else….*

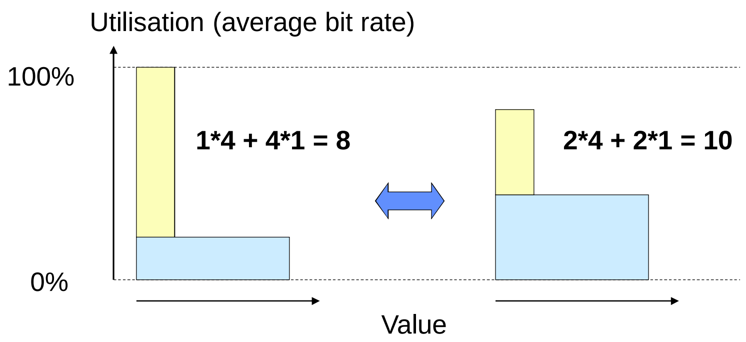
- Applications shall be considered through the business
  - particularly, the quality requirements of a flow shall be satisfied only if that is justified from business perspective
    - think, for instance, a situation in which someone is listening a streaming audio without paying anything

Customers´ role in the framework

- Customers are willing to pay (most probably) for
  - person to person communication
    - both voice and data (SMS!),
    - instead, video is doubtful
  - for real-time or personalized information (to some extent)
  - perhaps entertainment including games
- Extra for *mobility*, simple use etc.
- The viewpoint used here
  - the payment as such is important, rather than the reason why the customer is paying
    - in multi-service environment anything else is very hard to justify

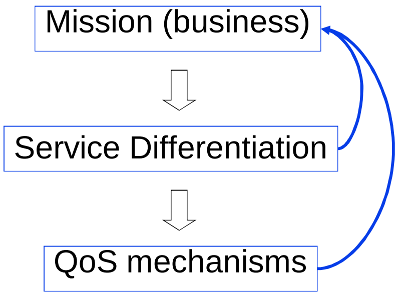
Throughput in the framework

- Total value (from business perspective) is the matter to be maximised
  - maximisation of throughput is relevant (as such) only if there is no value differences



Needed: Formal definition of target

- We need tools to define and assess
  - service differentiation
  - QoS rules and mechanisms
- In our framework the question is:



How to depict and model the effects of service differentiation and QoS mechanisms to the mission of the service provider?

Session 1, part B

Tools for evaluation

Key concept: Utility

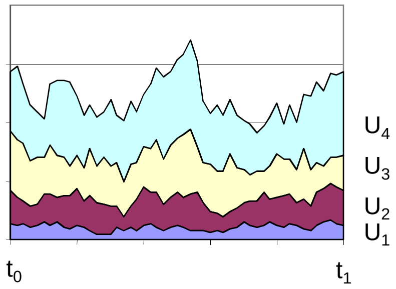
" the quality or condition of being useful "

- Utility of a flow of packets (U_i) depends on
  - service offered by the service provider
  - technical characteristics of the application
  - customer's preferences

Utility can be measured e.g., in money, wealth of society, or end-user's satisfaction

- Essential factor is the total utility
  - the sum of individual utilities measured over a period of time

$$U_{tot} = \int_{t_0}^{t_1} \sum_i U_i(t) dt$$



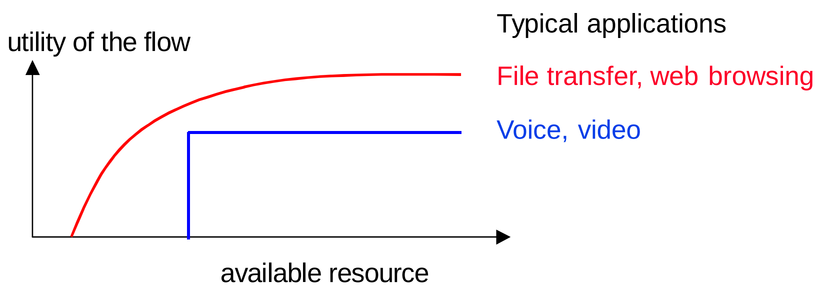
Utility functions

- *Utility function* describes the utility of a flow or an application as function of a quantity or quality parameter(s),
  - e.g., bandwidth, delay, bit error rate

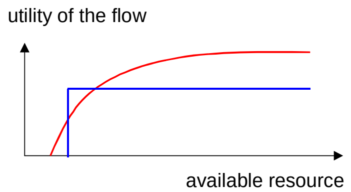
- Three issues to be considered
  - form of the function
  - horizontal scale
  - vertical scale

Forms of utility functions

- Primary forms
  - step: utility is either 0 (or even negative) or maximal
  - convex: the higher the available resource the better
- Form depends mainly on the application



Dimensions of utility functions



- Horizontal axis
  - bit rate
  - delay, bit error rate, etc
- Vertical axis
  - in general, usefulness of the application
    - expressed by the willingness to pay
    - defined by the customer (not application)
- Multi-service network
  - huge differences in both dimensions
    - bit rates from few kbit/s (voice) to several Mbit/s (video)
    - willingness to pay from 0.001 to 1000 $/Mbyte (!?)

Value differences

	Volume or bit rate	Acceptable price for transmission	Value ($/Mbyte)
SMS	160 bytes	0.16 $/message	1000
Voice	16 kbit/s	0.12 $/min	1
Video	2 Mbit/s	0.90 $/h	0.001

Good or bad news?

The whole point of service differentiation is to exploit these value differences;

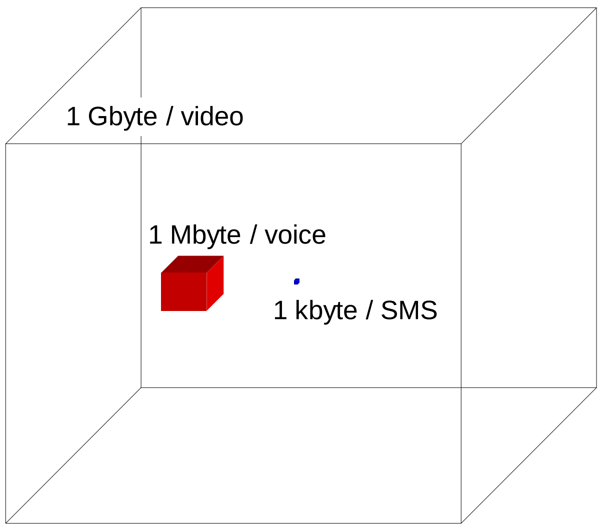
but pricing structure is problematic

Illustration of value differences

Needed amount of bytes for equal value
(3-dimensional)

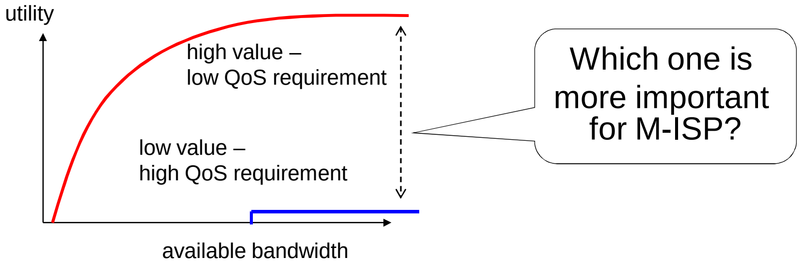
transmission time over a 8 kbit/s link:

1 s
17 min
12 days



Form vs. dimensions of utility function

- Main aspects
  - form of utility function defined by application = QoS requirement
  - horizontal scale, in some cases customer has an effect (Web!)
  - vertical scale defined (mainly) by the customer



Differentiation from utility viewpoint

- Goal of differentiation is maximization of total utility

$$U_{tot}(t) = \sum_i U_i(W_i, Q_i)$$

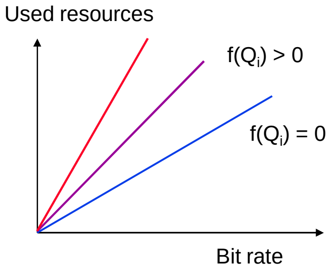
by giving different quantity (W_i) and quality (Q_i) of service to different flows

  - W_i = bandwidth, Q_i = {delay, bit error rate, ..},

- Constraint $\sum_i W_i \left(1 + f(Q_i)\right) \leq C_{tot}$

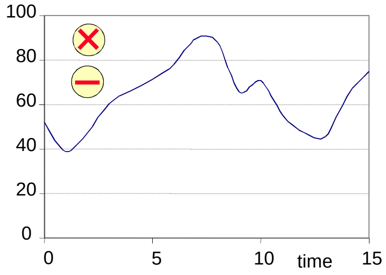
Quality Factor $(1+f(Q_i))$

- Specific quality requirements (delay, bit error rate) effectively means an addition to the used resources
- Delay requirement
  - because of small buffers utilisation has to be lower ⇒ effectively more resources are needed
  - for voice: 1+f(Q) ≈ 1.2 (smooth traffic) … 2 (bursty traffic)
- Bit error rate
  - particularly in radio networks the coding scheme effects on the bit error rate
  - e.g., in GPRS 4 coding schemes from 9 to 21.4 kbit/s



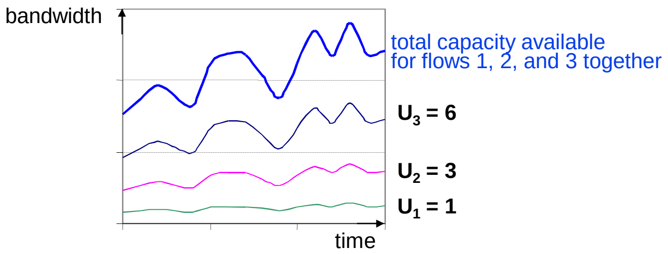
Services: Step functions

- Service differentiation with step-like utility
  - maximization of total utility means (roughly) that
    - the system shall accept entire calls with the highest revenue per used resources (U_i/W_i)
- Efficient solution would require that capacity variations and service requests could be accurately predicted
  - in practice, acceptance decision may depend on *U/W* of the flow and on the load level
  - calls with low *U/W* are
    - terminated
    - not accepted



Services: Convex functions

- Service differentiation with convex utility
  - maximization of total utility means (roughly) that
    - the system shall divide the total resources proportional to the price paid by the customer $W_i = c \cdot U_i$
    - (quality requirements can usually be ignored)



Combination of step and convex

- Only reservation-based service
  - hardly applicable with data applications, like web browsing
- Only sharing of resources
  - sufficient quality for critical applications (from business perspective) shall be guaranteed
- Division into two parts
  - one part for applications with step-like utility function
    - admission of "calls"
  - another part for applications with convex utility function
    - sharing of resources
  - but how to divide the capacity between the parts?
    - once again, the division shall serve the business!

Example

- Flows
  - voice: 25 flows, 20 kbit/s, real-time, P_{loss} < 1%, 0.1 $/min
  - video: 1 flow, 250 kbit/s, near real-time, P_{loss}< 0.001%, 1 $/h
  - web: 50 flows, 10 kbit/s (average), adaptive, 0.5 $/h
- What flows should the ISP favour during congestion?
  - video because of highest quality requirement?
  - web because of lowest quantity requirement?
- Utility per used resources
  - voice: U/W = 0.67 $/MB
  - video: U/W = 0.01 $/MB
  - web: U/W = 0.11 $/MB

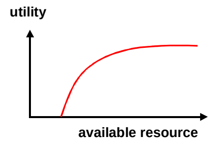
Importance:
voice > web > video

Summary of Session 1 Objectives

- Essential things are often simple
  - the difficulty is to keep them in mind and apply them in real situations particularly when some traditions are against them
- Here
  - total revenue is the matter to be maximised
    - not throughput
  - the value of a flow is defined by the customer
    - there is no direct relationship between application and the value of the flow
  - quality requirements of applications are of secondary importance
    - delay in case of voice is an important quality requirement

Summary of Session 1
Utility - The Tool

- *Utility function* captures the most essential characteristics of an application
  - quantity requirement (bit rate)
  - quality requirement (form of the function)
  - utility (user's readiness to pay)



- Total utility provides an unambiguous definition for the goal of service provision

- Together these can be used to evaluate any kind of service provision system

End of Session 1

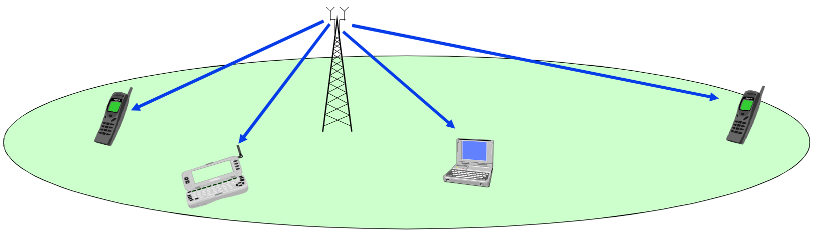
start	duration	topic
13:30	0:10	Introduction
13:40	0:50	Objectives and tools
14:30	0:20	Coffee
14:50	0:20	Radio networks
15:10	0:40	GPRS, UMTS, WLAN
15:50	0:10	Break
16:00	0:40	Case study
16:40	0:20	Conclusion & discussion
17:00		The End

Session 2, Part A

General Aspects of Radio Networks

Technical system

- Radio access network
  - with centralized control (one base station)
  - with multi-service capability
- GPRS / UMTS / WLAN
- Mainly
  - technical issues from service differentiation viewpoint



Technical and business viewpoints

- Various technical features related to bandwidth, delays, coverage area, power, cost structure, etc.
  - effects are sometimes conflicting
- Technical view
  - power consumption of terminals is a critical issue
    - ⇒ limits the use of high bit rates and high quality displays
    - ⇒ may decrease the need of differentiation
    - as in the current Internet where the access bottleneck (< 56 kbit/s) restricts users from robbing bandwidth
- Business view
  - radio bandwidth is an expensive and/or scarce resource
    - ⇒ differentiation is more important than in fixed network

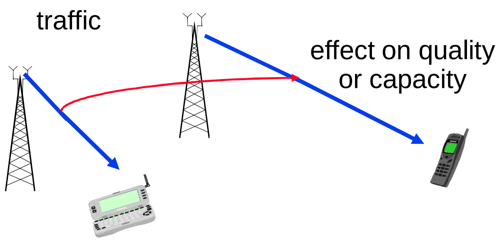
Radio-specific problems

- Unreliability of radio link
  - bit errors
  - fading
  - interference with other links
- Variability of radio link capacity
  - depends on distance and obstacles
- Delays
- Low bit rates
- Control of uplink capacity
  - shared resource in a complicated way

Bit errors

- Special characteristics of radio channel (multipropagation, fading) ⇒
  - Bit Error Rate (BER) on wireless link is much higher than in fixed networks
    - even 10^{-2}, during short periods practically 50%
    - can be solved at the expense of capacity and delay
- *Differentiation* aspects
  - trade-off between capacity and BER depends on application
    - with voice it might be better to have BER of 10^{-3} with 16 kbit/s coding than BER of 10^{-6} with 8 kbit/s coding
    - with data any bit error brings about re-transmission
  - possibility for some optimisation based on application
    - parameter available in many specifications

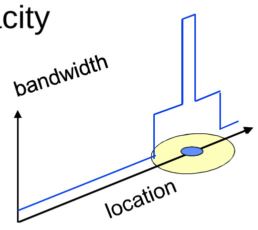
Interference



- Particularly a problem in ad-hoc WLAN without centralised control
- From service differentiation viewpoint
  - maximisation of throughput or total utility is not a local issue
  - complicates the implementation of service differentiation
  - sending of "useless" traffic shall be avoided

Variability of (effective) capacity

- Dependency between range and capacity
  - within a picocell ⇒ very high bit rate
    within a microcell ⇒ high bit rate
    macrocell ⇒ low bit rate



- This (together with fading, interference etc.) means that the available capacity between base station and mobile terminal is highly variable and difficult to predict
- From service differentiation viewpoint
  - general guarantees are almost impossible ⇒ sharing of resources or relative priorities are more practical
  - fixed price is very problematic (e.g., 1 $/h for 200 kbit/s video)

Delays

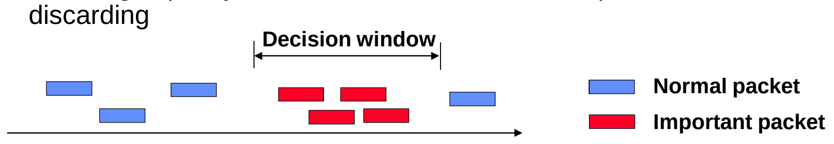
- Propagation delay
  - round trip time, 100 m: 0.6 µs, 1 km: 6 µs
    - not as such a significant problem
  - delay spread (due to multipath propagation) is more problematic
- Radio turn-around time
  - to switch from transmission to reception mode
  - typically 5 µs to 30 µs
  - increases the length of radio frame ⇒ adds delay variation
- Handling of BER-problem induces additional delay
  - ARQ (Automatic Repeat Request), interleaving
- Additional delay complicates service differentiation

Low bit rates

- Total bandwidth is much smaller than in fixed network
  - transmission delay of even one packet can be significant

1 kbyte packet & 1 Mbit/s link ⇒ 8 ms transmission delay

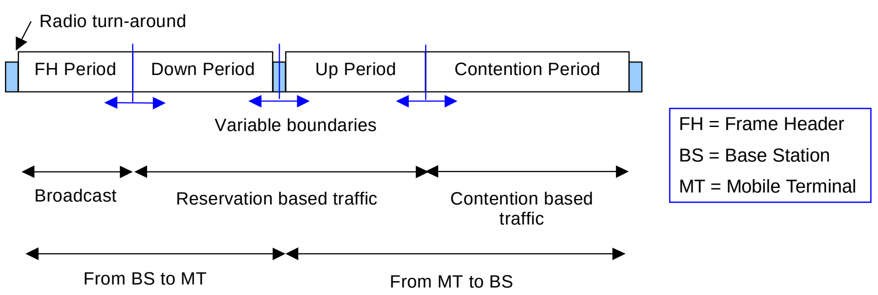
- From service differentiation viewpoint
  - buffering capacity is small ⇒ fewer choices for packet discarding



  - few flows ⇒ statistical multiplexing is inefficient ⇒ quality or quantity guarantees become difficult to implement and expensive

Capacity control

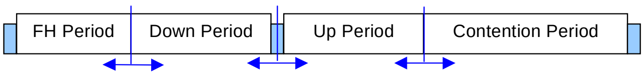
- Media Access Control (MAC)
- Design goal: efficiency and possibility for differentiation
- Example
  - MASCARA by Magic WAND Project



Controlling of shared radio resource

- Key issues from service *differentiation* viewpoint
  - (1) controlling of variable boundaries
  - (2) sharing of resources within each period (down, up, contention)
- Down (BT ⇒ MT) quite the same as fixed link
  - however, capacity is shared with uplink direction
  - granularity is not a packet, but rather a radio frame
- Up (MT ⇒ BT) the additional problem is that BT has only inaccurate information about transmission needs
  - requires signaling between MTs and BT ⇒ overhead
  - queues are distributed ⇒ fair differentiation difficult

Variable boundaries



- Frame Header
  - from efficiency viewpoint size of FH shall be minimized, but still enough information must be transmitted
- Downlink vs. Uplink periods
  - because Uplink and Downlink share the same total capacity, service differentiation shall concern both directions

    ⇒ should be as flexible as possible
- Reserved vs. Contention
  - reservations are more efficient with applications with constant bit rate, but problematic with bursty traffic

Session 2, part B

GPRS, UMTS, WLANs

General Packet Radio Service (GPRS)

- Mission
  - reuse of existing GSM infrastructure to provide efficient service for bursty packet-based applications
- Main characteristics compared to GSM
  - always on
    - much faster set-up and access time (e.g., for WAP)
  - customer may pay for transmitted information rather than the time spent on line
  - higher bit rates
    - (1 ... 8 timeslots) * (9 … 21 kbit/s)
    - in practice, maybe 50 kbit/s depending of circumstances (BER, load, handset properties)

GPRS: Service structure

- Attributes for a flow (or PDP context)
  - Precedence class (3 choices)
    - high, normal, low; used to determine which packets shall be discarded during congestion
  - Delay class (4)
    - delay requirement, from 0.5 s upwards
      not for true real-time applications
  - Reliability class (5)
    - error sensitivity related to bit errors, data loss etc.
  - Throughput class (9 + 19)
    - expected maximum bit rate; 8 kbit/s … 2 Mbit/s
    - average bit rate; best effort to 111 kbit/s (over 1 hour)

GPRS: Differentiation vs. classes

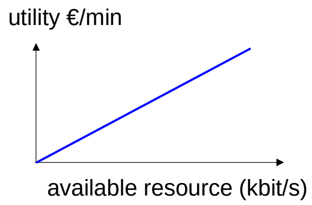
- Precedence class
  - might be usable if bound to pricing
- Delay class
  - seems to be more about the short-term division of resources than actual delay through the network
  - limited usability from differentiation viewpoint
- Reliability class
  - optimal selection seems to depend more on characteristics of the radio link than the application
- Throughput class
  - if charging per byte, then the whole capacity should be used as efficiently as possible, not based on bit rates

Utility viewpoint

- Analysis from service differentiation viewpoint
  - 3*4*5*9*19 = a large number
    - (though many combinations are not relevant)
  - requires simplification in real implementations
- If price is the same constant per byte,
  - then what is the use of quality parameters?
  - actually, it is assumed that

  ⇒ no reason for differentiation within GPRS

  ⇒ the main question is about differentiation between data and voice



Universal Mobile Telecommunications Service (UMTS)

- Mission
  - provision of universal service access and mobility support, including both personal and terminal mobility
- Main advantages – typically mentioned
  - more capacity
  - Quality of Service

  ⇒ possibility to offer video and other attractive services

UMTS: Service classes

Note: the meaning of UMTS classes is different than GRPS classes

- Conversational
  - for voice (or for any application with similar delay requirements)
  - provides low delay (of the order of 100 ms or less)
- Streaming
  - for one-way transport of video or audio streams
  - delay of some hundreds of milliseconds
- Interactive
  - for web browsing, or similar interactive application
- Background
  - file transfer, e-mail, without any specific time constraint

UMTS: QoS parameters

Note: the meaning of these parameters is similar to GPRS classes!

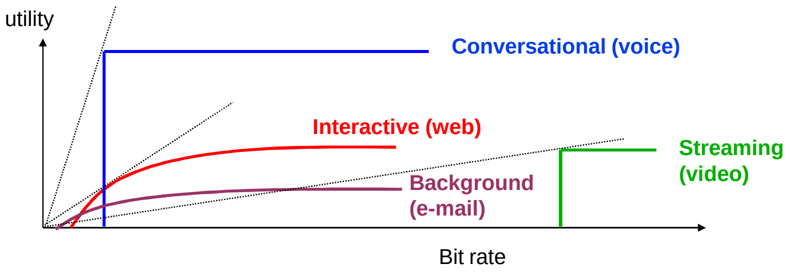
- Traffic handling priority
  - only in interactive class
- Transfer delay
  - somewhat overlapping with service classes
- Residual BER & SDU (Service Data Unit) error ratio
  - 3…7 levels depending on service class
- Maximum bit rate
- Guaranteed bit rate
  - only in conversational and streaming classes, up to 2 Mbit/s
- etc.

UMTS: Service differentiation viewpoint

- Remarks (personal)
  - QoS system has too many parameters to be used in its entirety
  - not applicable without a clearer framework
- What is essential from business viewpoint?
  - Special delay (or service) class is needed for voice, others unclear
    - 3 classes are likely enough
  - Prioritisation based on business importance is needed
    - cannot be based on applications requirement
    - should cover all service (or delay) classes
  - Bit rates
    - perhaps yes, but closely related to the previous item

UMTS: Utility viewpoint

- Business importance vs. UMTS framework
  - how to use UMTS system to optimise business – this is the critical issue!
- Starting point is the differences in utility and bit rate



UMTS: Utility consequences

- Voice shall typically have a high priority
  - but this must be tightly connected to price and bit rate
    - if the price paid by the customer is low or bit rate high, there is no reason to give high priority
- Interactive and background
  - belong to the same category
  - capacity (left by voice) should be divided proportionally to the price paid by each customer (evenly if the same price)
- Streaming (video or audio)
  - if and when the readiness to pay per byte is much lower than with other services, streaming applications are served only when there is free capacity

Wireless LANs (WLANs)

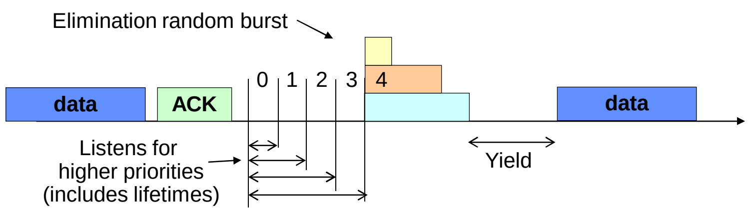
- Mission
  - to provide efficient and inexpensive radio access within a local area
- WLAN specifications are more about MAC than services
  - concentrate on throughput and some QoS features
- Examples
  - IEEE 802.11
  - HIPERLAN/1
  - HIPERLAN/2 (Wireless ATM)

IEEE 802.11

- = MAC protocol
  - for different physical layers, for centralised and ad-hoc networks
  - up to 20 Mbit/s, typically 1 - 2 Mbit/s
- Basic channel access does not support any QoS
  - based on CSMA/CD (Carrier Sense Multiple Access/Collision Detection)
  - very inefficient with short packets
    - total overhead 300% with 40-byte packet at 2 Mbit/s
- PCF (Point Coordination Function)
  - supports contention-free service
  - optional, but apparently not used in reality

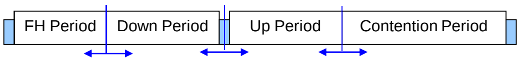
HIPERLAN/1

- = Architecture for local ad-hoc radio network
  - 5 GHz band (Europe)
  - from 1 Mbit/s (800 m) to 23.5 Mbit/s (50 m)
  - supports both asynchronous and time-bounded services
- No central controller ⇒
  - no controlled service differentiation
  - nevertheless, 2 user priorities, and residual lifetime



HIPERLAN/2 (Wireless ATM)

- = Wireless access for managed infrastructure
  - centrally controlled with reservation slots (= ATM cell)
  - user data rates up to 54 Mbit/s (at least in theory)
  - MAC = Mascara:



- MAC Virtual channels provides QoS features
  - supports different delay and reliability levels, and bandwidth reservations

  ⇒ also suitable for voice and video

Service differentiation in WLANs

- Availability of QoS mechanisms is not the same as controlled service differentiation - quite the reverse
  - prioritisation is dangerous if not controlled appropriately
  - IP packets have contained Type of Service field (drop precedence, delay, reliability) for years with very limited use
- Without service provision infrastructure
  - service should be based on best effort principle
    - with as fair resource allocation as possible
    - possibly, an additional support for real time applications, but in a way that they do not rob bandwidth
- With service provision infrastructure
  - most important issue is to design appropriate connection between charging and the quantity and quality of service

Summary of Session 2
Radio-specific issues

- Radio access is
  - on the one hand, expensive resource with peculiar characteristics
    - ⇒ service differentiation might be worth of effort
  - but on the other hand, unreliable shared resource
    - ⇒ makes hard guarantees almost impossible
- From service viewpoint
  - too ambitious objectives, e.g. absolute guarantees, are risky
  - effort shall be used to develop most essential things that have significant effect on the service provider business
    - including implementation and management cost

Summary of Session 2
UMTS, GPRS, WLAN

- UMTS
  - complex system with classes and parameters
    - requires a clarified framework
  - ISP business must be the starting point
  - utility calculations is a tool to design service differentiation
- GPRS, WLAN
  - most suitable for best effort, data services
  - WLAN may have limited support for real-time applications
    - any other service differentiation is reasonable only with appropriate relationship with charging

End of Session 2

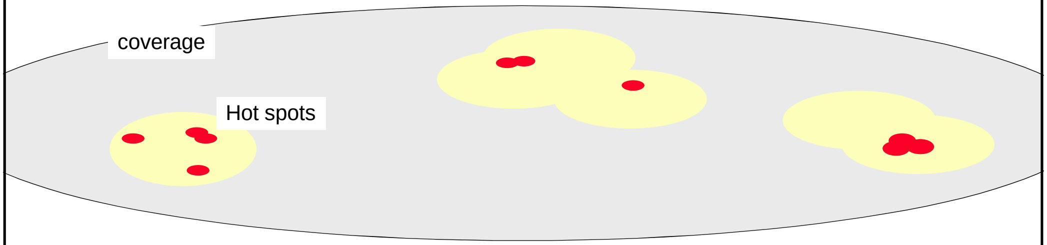
start	duration	topic
13:30	0:10	Introduction
13:40	0:50	Objectives and tools
14:30	0:20	Coffee
14:50	0:20	Radio networks
15:10	0:40	GPRS, UMTS, WLAN
15:50	0:10	Break
16:00	0:40	Case study
16:40	0:20	Conclusion & discussion
17:00		The End

Session 3, part A

A case study

Case study: Infrastructure

- Service provider with a cellular network
  - wide area coverage with relatively low bandwidth
    - average bit rate per available for a customer = 30 kbit/s
  - hot spots with essentially higher bandwidth
    - can be done with any technology, e.g., WLAN on unlicensed band
    - average bit rate per available for a customer = 200 kbit/s



Services (1)

- Conversational = voice
  - quality: as good or better than in GSM networks
  - coverage: shall work everywhere
  - charging: 0.1 $/min minute
  - main target: to ensure that the profitability of current voice business continues
- Interactive (web, wap)
  - quality: good enough latency most of the time
    - but depends strongly on available bandwidth
  - charging: primarily flat rate = 20 $/month
    - services can have additional charging components not directly related to transmission of bytes
  - main target: to expand service offering and attract new customers

Services (2)

- Background (e-mail, photographs, etc.)
  - quality: reasonable downloading time;
    applications and customers have to adapt to the reality
  - charging: flat rate = 20 $/month
  - main target: to make it possible to expand the variety of applications used by mobile terminals
- Streaming (video, audio)
  - quality: general quality guarantees are infeasible, but within hot spots some expectations can be created
  - charging: basic service with 20 $/month
    additional quality with 0.1 $/min
  - main target: to create favourable image about the services provided by the ISP

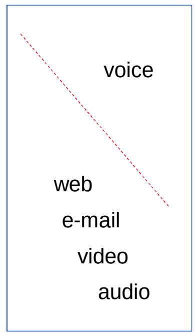
Main principles of differentiation

- Whether a flow is significant or insignificant is up to the customer, and that does not depend (much) on the application or service class
  - the same application can be of much higher importance for one customer than for another one
  - customers express the importance of flows by paying

  ⇒ starting point of designing QoS is charging,
  not classification of flows based on applications
- Other aspects to be addressed
  - applications may have certain importance explicit payment
    - advertising value
    - general reputation
    - potential to generate income in future

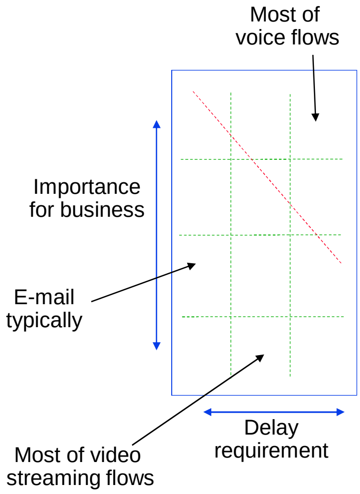
Service differentiation

- Two main categories: voice and "flat rate"
- Voice in its own category because of
  - business importance
  - charging principle
  - step-like utility function
  - delay requirements
- Everything else
  - interactive, streaming, background
- in the other category
  - based on flat-rate charging and flexible sharing of resources



Sub-categories

- Classification can further be elaborated by introducing a sufficient number of delay and importance classes
  - the place of each flow or packet in this framework depends on
    - price, and possibly other value related to the flow
    - bit rate, and other resources used by the flow
    - delay requirement of the application



QoS mechanisms

- The objective of the QoS mechanisms is to realise the service differentiation based on sub-categories
  - importance and delay dimensions
- Most important flows (e.g., voice) attain highest priority and/or resource reservation, with excellent quality (including delay)
  - at the expense of time-dependent price and limited bit rate
- All other flows share the remaining capacity in a controlled manner
  - each flow gets approximately the same bit rate, but the bit rate depends heavily on load situation and the capacity of the cell
  - higher share can be obtained by paying extra

Radio-specific issues

- If a flow or application requires special quality (BER, delay), the network shall of the characteristics,
  - but at the expense of bit rate
- Sharing shall concern both uplink and downlink at the same time
- Radio frame limits the dynamics of the sharing
  - sharing is not necessarily as "exact and fair" as what is possible in a fixed network
- Significant variations in available bit rate and quality
  - both applications and customers have to be adaptable

Final outcome

- Percent of active customers (during busy / idle hours)
  - voice: 60 / 40
  - web type of applications: 25 / 25
  - file transfer including e-mail: 10 / 20
  - streaming video: 5 / 15 (note: requires extra fee during busy hour)

- Outcome bit rate kbit/s

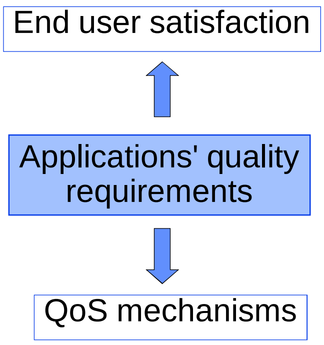
  FR = flat rate
  extra = 0.1 $/min

Coverage area	Broad	Broad	Hot spot	Hot spot
Busy/idle hour	Busy	Idle	Busy	Idle
Voice	enough	enough	enough	enough
Web, E-mail	30	60	200	500
Streaming, FR	not enough		possibly enough	likely enough
Streaming, extra	100	200	500	1000

Session 3, part B

Final Remarks

Traditional QoS model



It is assumed that customers are happy if and only if applications´ requirements are satisfied

everything seems to start from the applications´ requirements

QoS mechanisms are derived from those requirements

Why the traditional model is so attractive?

- The traditional model to consider QoS is, and will be popular (perhaps, forever)
  - still, we have to be aware of its weaknesses
- The next three slides provides some, hopefully fresh, viewpoints to re-consider the basis of QoS
  - mainly for those that need a comprehensive understanding about QoS,
  - but everyone is entitled and encouraged to form his or her own opinion, and free to move directly to slide 76

Viewpoint 1: Perfect QoS System

Let us consider a set of applications and the related flows. We are inclined to think that if the requirements of an application are not satisfied, the QoS system is imperfect. Since we are experts of QoS, our goal is to design a perfect system, and we must change the QoS system in a way that the requirements will be satisfied. Once that requirement is met we can proceed to the requirements of next application, and make some new changes or additions to the QoS system. Finally, when all requirements of all applications are satisfied, the system is perfect and ready. The underlying assumption is that a perfect system really is possible, and that our task just is to design and build it; in the same way as physicists are developing the grand unified theory. Although we may be somewhat troubled with the complexity of the system, the process seems to guarantee the right outcome, if we just were patient enough.

But now there is a fundamental difference between the realms of physics and QoS systems. In physics, because we may assume that the world under consideration is consistent, it seems possible to incorporate all known facts into a single model, at least in theory. In contrast, there is absolutely no guarantee that all the QoS *requirement*s which the QoS system must cope with were consistent. As a naive example, customer A may require that the quality of his service is better than that of customer B, and customer B may require that the quality of his service is better than that of customer A. In real cases, the inconsistency is much more concealed but still very real.

Viewpoint 2: Taxation principles

We may contrast the desired treatment of IP packets with the desired treatment of individuals. What seems to be a reasonable decision from the perspective of an individual citizen (e.g., removing of fuel tax, not perhaps entirely, but at least in case of important customers like me), is not necessarily reasonable from the perspective of the whole society. In the same way as taxation is a matter of the whole society, the rules to decide which IP packets shall be discarded is matter of the wholeness of service provision.

Analogically, we may imagine a fuel tax system in which the tax is based on the requirement of each customer filling the tank of his car: an aggressive motorist requiring low tax will get lower tax, or perhaps, the tax will depend on the model of the car. Of course, from a society viewpoint these taxation principles are senseless.

The feeling we get when thinking a situation from the viewpoint of an individual person is often very strong; it is fair and right that the person gets what she wish and expects. Unfortunately, the sum of the results desired by a large number of individual persons do not form any coherent system. Similarly, we cannot design any coherent QoS system based on the treatment desired by individual packets despite the strength of feeling evoked by specific situations.

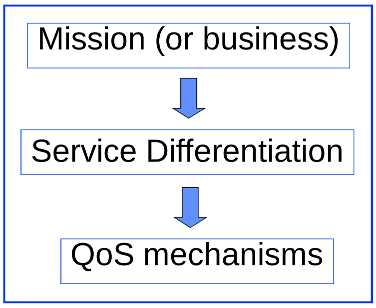
Viewpoint 3: Modes of thinking

Apparently we have several modes of thinking, and one of the modes is dominant when thinking a specific matter, like QoS. If someone is a technical expert and the matter is at least partly technical, the mode of thinking most likely is technical as well. As a result, the whole issue is considered from technical viewpoint; for instance, the task of QoS mechanisms appears to be to satisfy the technical requirements of applications.

Then if the same person is discussing about the prospects of UMTS services at coffee table, the viewpoint can be different, e.g., the viewpoint of real customer or of a service provider. Still it is very difficult to convey even very clear and relevant observations and conclusions from those areas into the realm of technology. Transferring of an observation from one area into another one requires the change of thinking mode, and that change also means a change of concepts, a change of objectives, and even a change of world of ideas.

What appeared as a very strong conclusion a minute ago when discussed about the prospects of UMTS operators, is somehow magically weakened when moved into technical area of QoS mechanisms. We have to satisfy all the technical requirements of an application regardless of how nugatory we consider the application, haven't we.

Service differentiation
is a tool used by service providers
to support their mission



Mission is the starting point and reference for all other matters,

service differentiation shall serve the mission,

QoS mechanisms shall serve the service differentiation

Final questions

- What exactly is the advantage of a service differentiation rule or QoS mechanism from the viewpoint of service provider's business?
- Does the value of the advantage exceed the cost of implementing and managing the rule or mechanism in real networks?
  - Ask these questions whenever someone proposes a new (or old) service differentiation rule or QoS mechanism
  - You may hint at utility function as a tool to formulate an appropriate reply

The End

∞

Thank you!