Internet services and applications

SLA, QoS

Internet-based services

- The internet for designed to support a wide range of services from the beginning.
  - *Provides „nothing*” but supports „everything”!*

    (= except connectivity!)*

- This goal was achieved by separating *service intelligence* from *data transfer*.

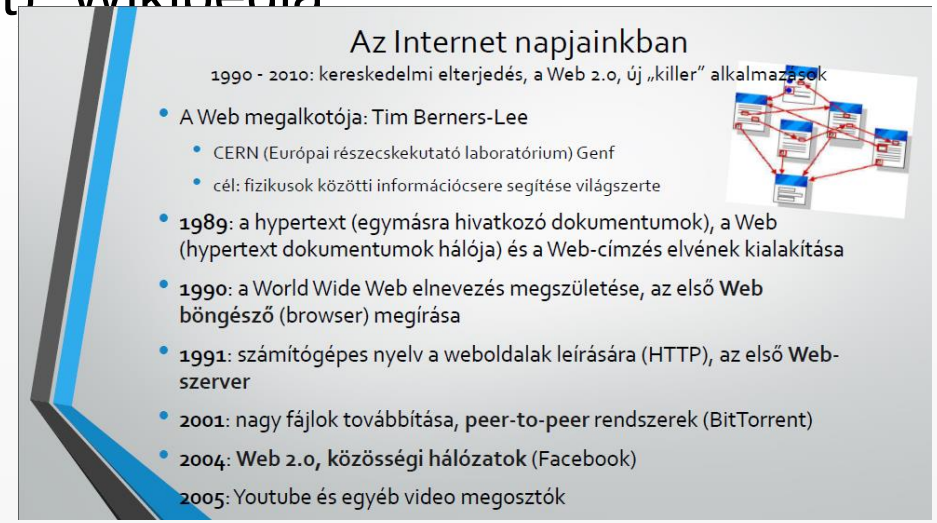
- The use of internet as a general purpose service platform became possible, when a special overlay, the World Wide Web, and its client application, the web browser appeared.

World Wide Web as a service platform

- Before the era of Web and web browsers, the service providers had to develop and manage their own end-to-end service resources.
- From 1990s the Web became th
  internet.
  - The Web provides an *universal s*
    Transfer Protocol);
  - and a *client application* in the forr
- Consequences:
  - The Web provides a fast time-to-r
  - reduces the learning time of end
    user interface for the service.

The internet not too long ago...

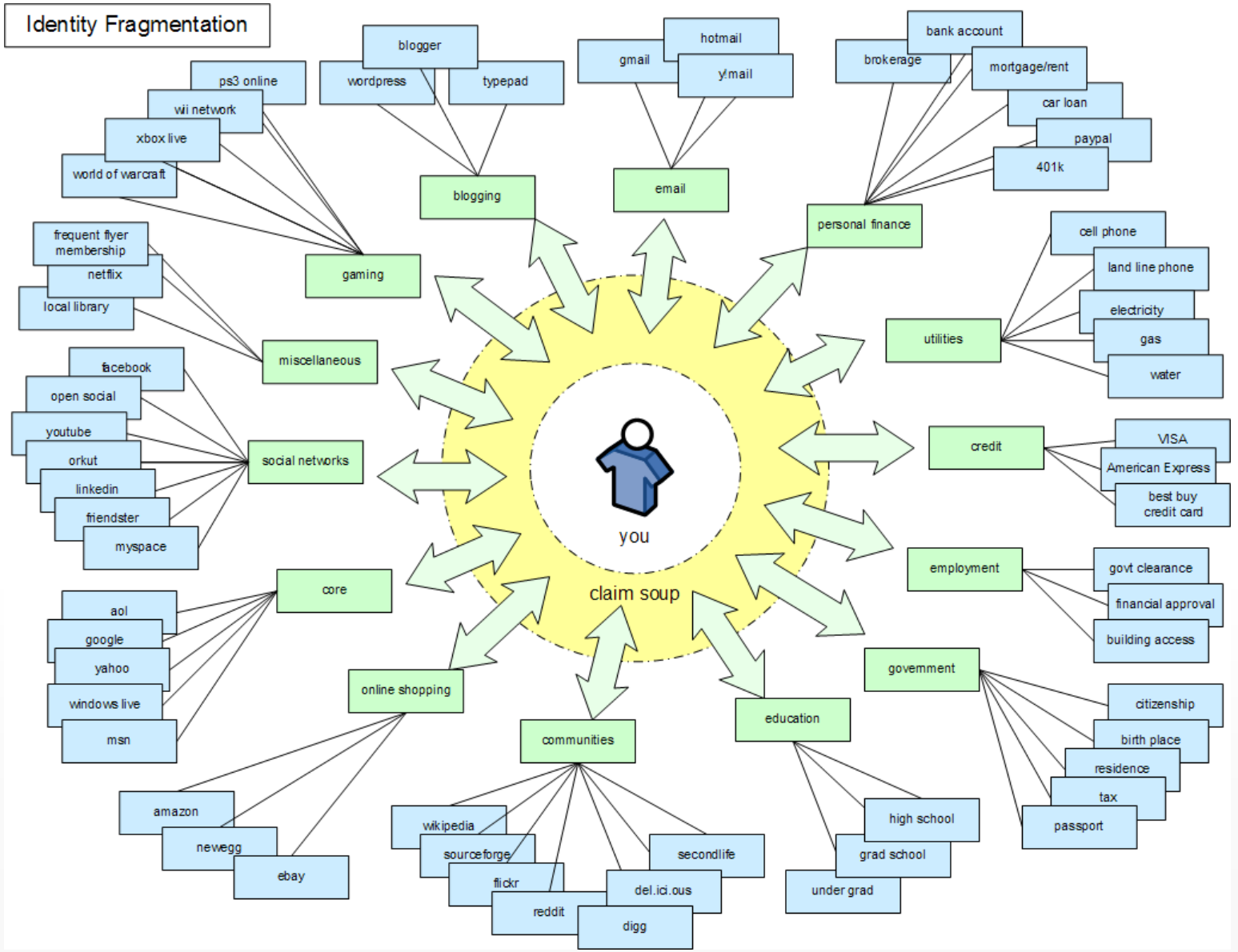
- 1989: hypertext
- 1990: WWW naming, first web browser appears
- 1991: HTTP, first web server established
- 1995-96: Amazon and eBay
- 1998: Google search engine
- 2001: peer-to-peer overlays (BitTorrent), Wikipedia
- 2003: Skype
- 2004: Web 2.0, Facebook, Youtube
- 2005: Reddit
- ...



Reminder...

- Services provide **<u>abilities</u>** for the **end user**

→ the end **user is in the center**, <u>not</u> the network!



Content

- SLA – Service Level Agreement
- Quality of Service (QoS)
  - QoS attributes, human factors, QoS zones
  - QoS handling in abnormal situations

Internet and QoS

- The influence of internet technology on the communication sector is HUGE!!

- But what is the Internet?

technologycal view

„global IP network”,
or
„everything that is above IP *(e.g., email, WWW)*”

users’ view

- When we talk about „Internet services” we have to distinguish...
  - **transport** services (like IntServ, DiffServ, etc.),
  - **end-system** services (Web services, email, telephone services, etc.)

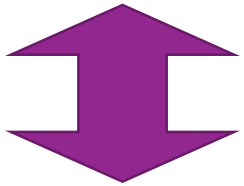
Service Level Agreement (SLA)

Content

- **SLA – Service Level Agreement**
- Quality of Service (QoS)
  - QoS attributes, human factors, QoS zones
  - QoS handling in abnormal situations

Service level agreement (SLA)

- Remark: Market liberalization and the development of services greatly influenced the definition of what we mean by "service"!
- "Previously" requirements for a service (e.g., telephony) were **defined by standards**, including centrally set tariffs.

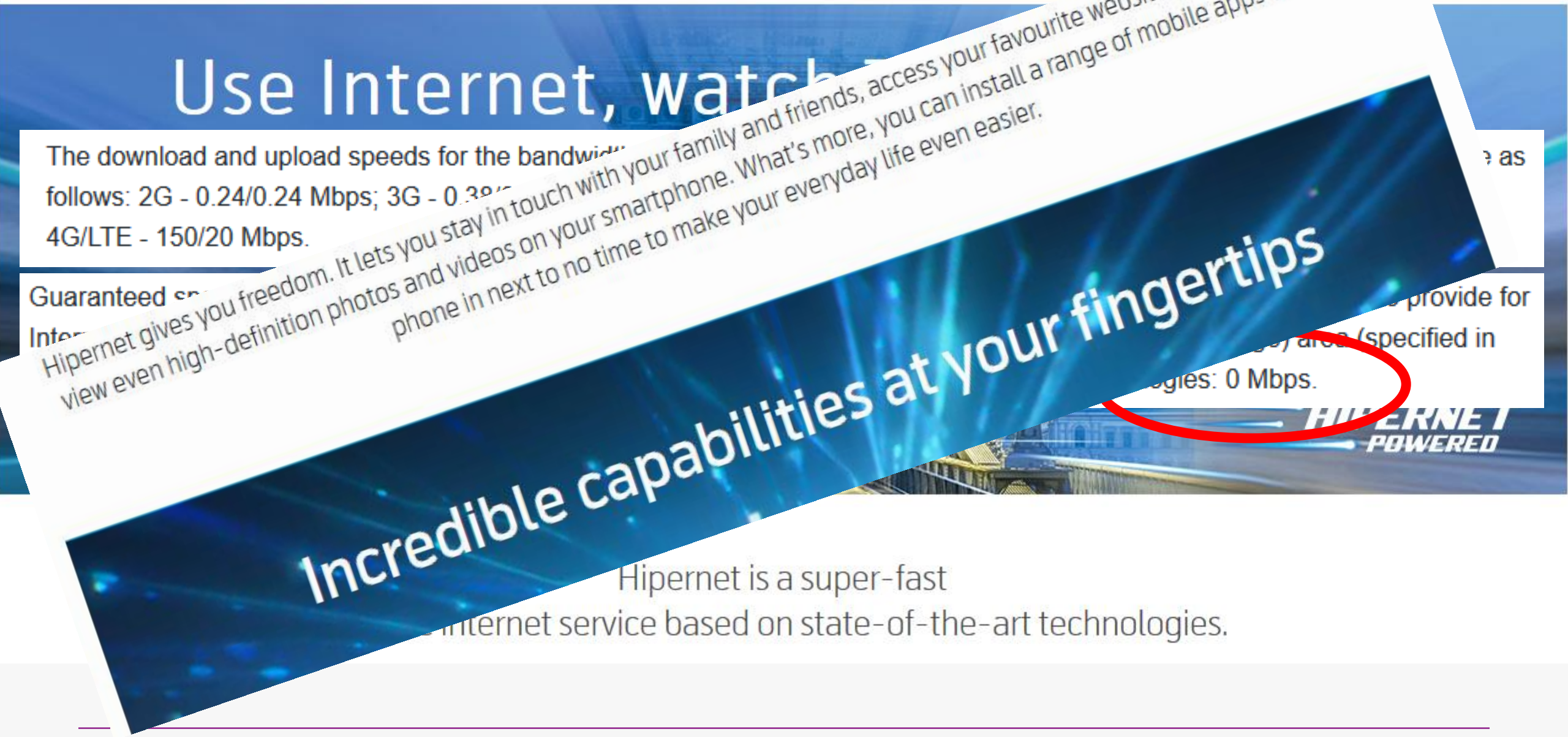


- Today, services are not strictly defined, so users and service providers specify requirements and tariffs in **contracts.**

- **SLA** = Service Level Agreement

Service level agreement (SLA) cont'd

- SLA-s specify…
  - service **limitations**,
  - the insured **quality of service** (QoS),
  - the **price** of the service.
- Eg, an SLA defines the ...
  - prices;
  - types of connections (eg voice, video, protocols);
  - quantity (eg number of channels or capacity [bit / s]);
  - reliability of data transfer (eg, tolerable bit error rate);
  - responsiveness (eg, connection setup time, server response time);
  - availability (eg, 24/7 with maximum annual x seconds outage)


Use Internet, watch
The download and upload speeds for the bandwi
follows: 2G - 0.24/0.24 Mbps; 3G - 0.38
4G/LTE - 150/20 Mbps.
as
Guaranteed sp
Inter
provide for
(specified in
gies: 0 Mbps.
Hipernet gives you freedom. It lets you stay in touch with your family and friends, access your favourite websites at any time and
view even high-definition photos and videos on your smartphone. What's more, you can install a range of mobile apps on your
phone in next to no time to make your everyday life even easier.
Incredible capabilities at your fingertips
HIPERNET
POWERED
Hipernet is a super-fast
internet service based on state-of-the-art technologies.

** The box contains at least 500 ml of milk. The guaranteed fat content is at least 1.4%. The information is not complete.*

Service level agreement (SLA) cont'd

- An **SLA** is a contract between...
  - provider and subscriber or ...
  - between two service providers.

- An SLA **between two service providers** can simplify service management tasks by partitioning the problem into separated regions.
  - This will avoid sharing the information needed to service management between service providers, but…
  - the price they pay for it is that the connection points between the service providers should be monitored to check compliance with the SLA.

Quality of Service (QoS)

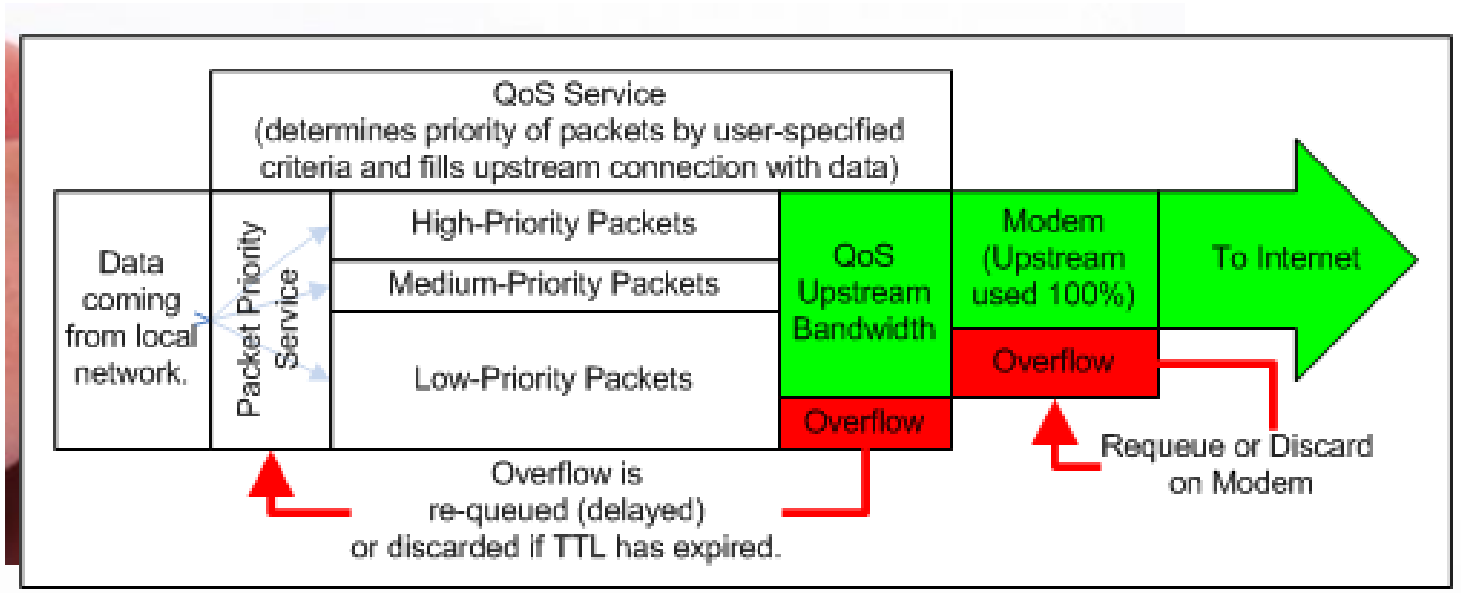
QoS attributes, human factors, QoE

Content

- SLA – Service Level Agreement
- **Quality of Service (QoS)**
  - **QoS attributes, human factors, QoS zones**
  - QoS handling in abnormal situations

Quality of Service (QoS)

- Note: In QoS, we mean **end-to-end QoS** for users that they experience



- (QoS could be specified at the internal points of a network or at the point of connection of two network service providers (eg within an SLA).
- Def: ***QoS can be defined with quality and timing features that are required to reach a level of service functionality and end-user satisfaction.***

QoS attributes

- Top quality attributes:
  - **Fidelity**: how accurate and faithful the source content was reproduced.
    - This is usually a function of available bandwidth, sampling resolution, and encoding.
  - **Loss:** lost packets of digital stream can result in lost details in source content
  - **Corruption**: Changing bits or packets could result in a change in source content.
  - **Security**: Ensuring that source content is not accessible to unauthorized customers.

QoS attributes (cont'd)

- Top **timing** attributes :
  - **Delay** (latency): the average elapsed time from sending the source data to the display at the recipient.
  - **Jitter**: The degree of deviation of the current delay from the average.
    - For example, jitter is a measure of how much the minimum and maximum (packet) delay deviates from the average for a media stream.
  - **Synchronization**: Delay difference for more than one media stream to be transmitted together (eg, audio and video).
  - **Set-up time**: how much time is needed for the service to build up
    - (Eg, telephone dial tone after lifting the handset)
  - **Tear-down time:** how long does it take to close access and release resources while a new set-up can be initiated.

Typical QoS compromises

- Improving one or more QoS parameters often affects other QoS attributes (if we do not assume extra bandwidth and processing capacity for it).
  - Eg, reduction of loss through available retransmission protocols, which usually increases the delay.
  - Eg, jitter and synchronization can be improved by buffering incoming streams, which also increases the delay.

- In general, the **delay** **seems to be the perpetual QoS attribute**.
  - Especially in IP networks!

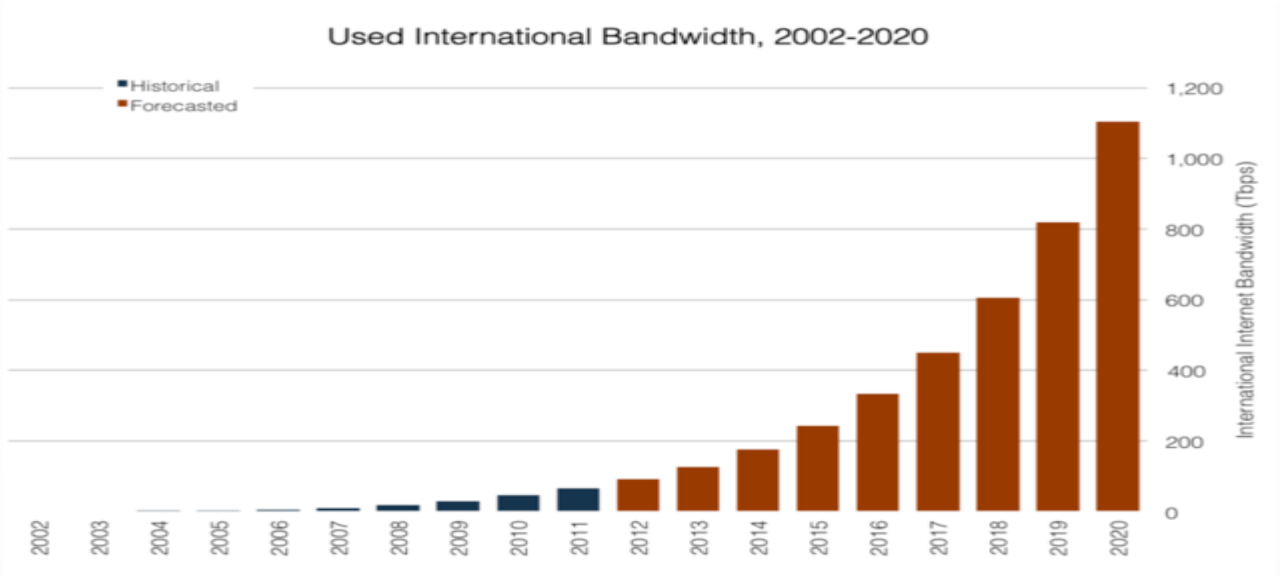
QoS attributes, human factors, QoS zones

Content

- SLA – Service Level Agreement
- Quality of Service (QoS)
  - QoS attributes, **human factors, QoS zones**
  - QoS handling in abnormal situations

Human factors and QoS

- Many times they say "QoS requirements are getting tighter as **users always want more and faster**"…



- But **human factors** - unlike technology - are well-understood and **unchanging** things.
  - For example, in video, the number of images needed per second (24-30) is sufficient due to the inability of the eye and brain imaging. This standard is already 100 years old!

Service Types depending on QoS

- Services can be classified into **four categories** based on the delay *if QoS is viewed from the human factors' point of view*:
  - This model was used to construct ITU-T Recommendation G.1010 (End-User Multimedia QoS Categories).
- **Perceptual**
  - The limits of the human sensory (eg hearing, vision) system limitations.
  - These barriers are the most stringent in terms of delay, typically less than 200 msec.
- **Cognitive**
  - Boundaries due to the limitations of short-term memory and the maintenance of natural attention, approx. 0.25 to 3 second intervals.

Service Types depending on QoS (cont'd)

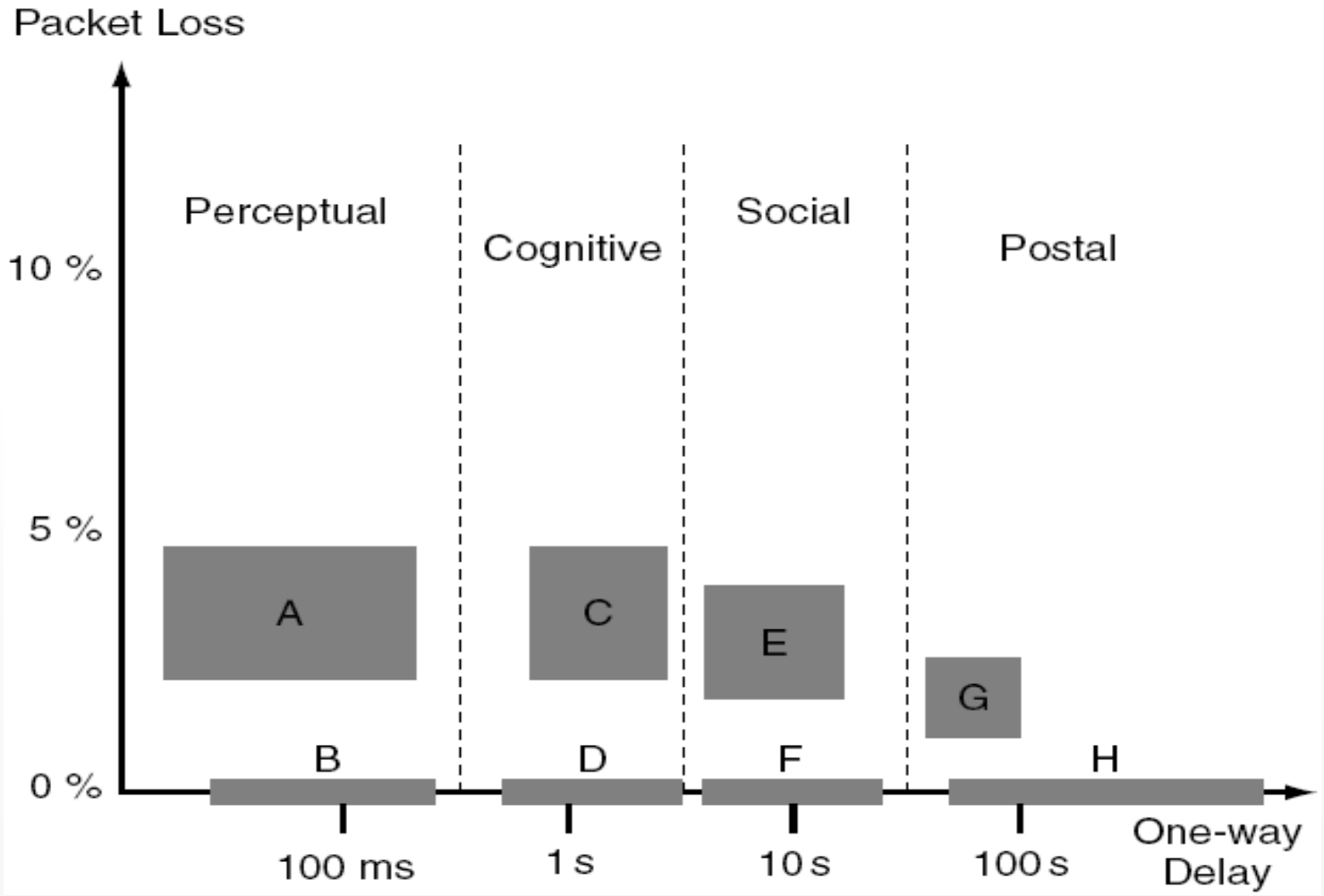
- **Four categories of services by delay (cont.)**
- **Social**
  - Borders on the basis of social expectations, expecting reasonable response times after a question or request.
  - Understanding of users can reduce these limits for a complex question / request.
  - Typical delays up to **10 seconds**.
- **Postal**
  - Expectations for delivering certain things (eg letter, fax) to another party.
  - The expectations are from 10 seconds to **several minutes**, possibly for **hours**.
  - Contrary to the social category, in this case the delay is not experienced by the outsider but by the receiving party. That is why the requirements are looser.

Service Types depending on QoS (cont'd)

- **Perceptual** and **cognitive** categories are neurological, so they are valid for every human being.
- **Social** and **postal** categories, on the other hand, depend on cultural expectations and experiences
- With human terminology, reproduction of source content must be **precise** (ie, error-free) or **forgiving** (ie, low error rate even without consequences)
  .
  - For example, the precise typically corresponds to a digital source (eg, bank account number), while the forgiving typically corresponds to an analogue source (eg, favorite color).

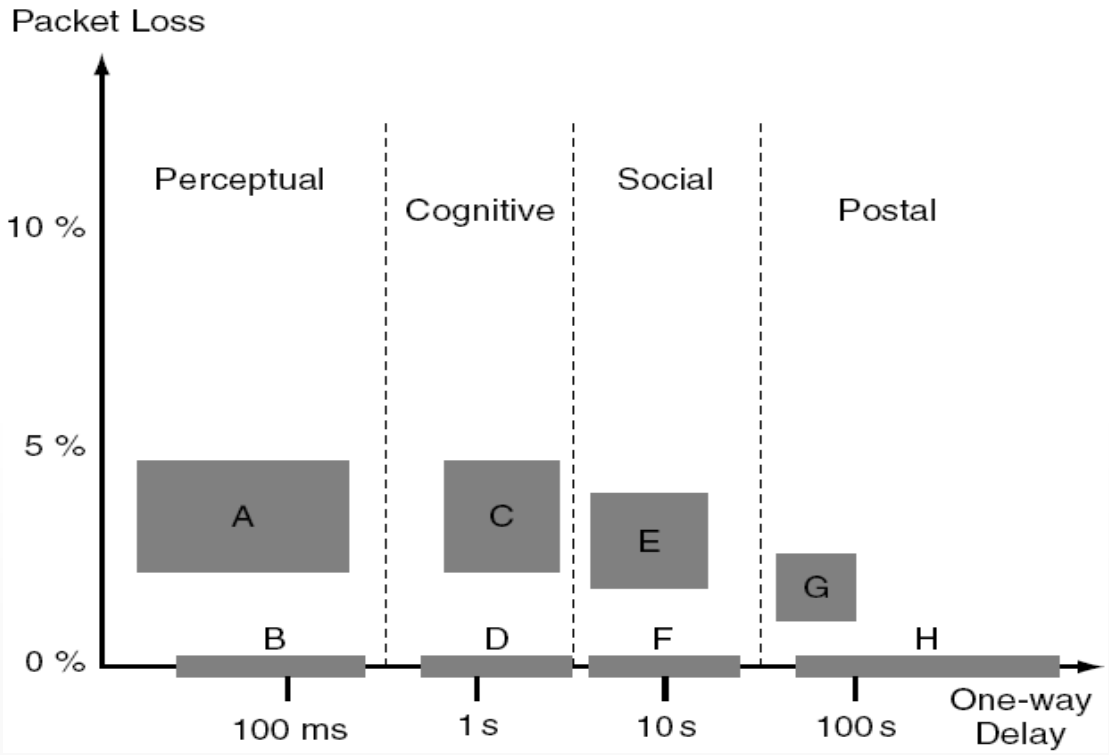
Service Types depending on QoS (cont'd)

- The four QoS delay categories, together with the two types of reproduction, designate **eight target zones**, taking into account the delay and loss:



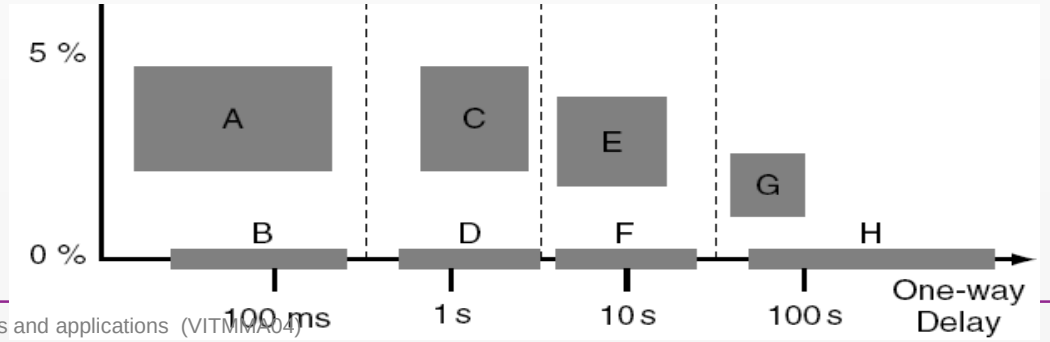
QoS target areas

- **A, C, E, and G target areas** for source media that endure certain losses (eg analog) and where loss can be traded with delay
- **Areas B, D, F and H** are target areas for source media that require 0% loss (eg digital), and where no compromise between loss and delay can be made. The only variable QoS attribute is the delay.



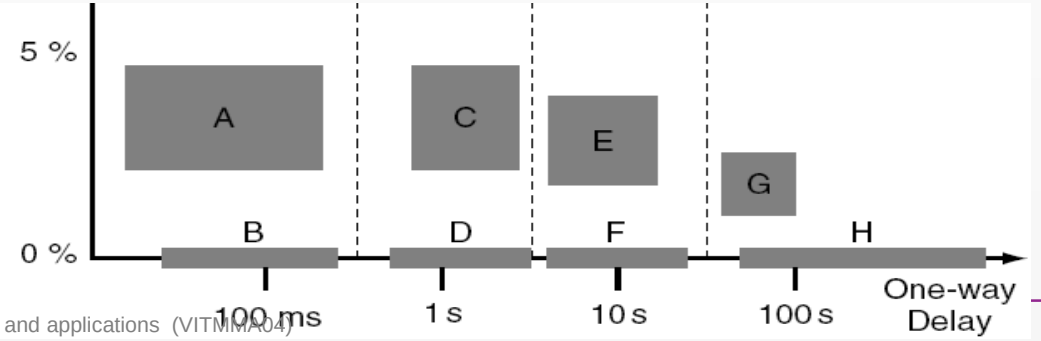
QoS target areas (cont'd)

- **Zone A: QoS for perceptual and forgiving media**
  - Typical service examples: **Two-way interactive** voice and / or video.
  - Due to the analog nature of the source content, it provides a continuous stream of information.
  - **Certain information loss can be tolerated** as the human hearing and vision system can compensate for different "noises".
  - The delay should typically be **below 200 msec**.
  - Delay exceeding the limit will result in delicate delay and / or synchronization sliding in the conversation (eg, through satellite transmission). These effects can cause emotional irritation and frustration, undermining trust between the parties.
    - Note: In some cases, the parties may eliminate these effects, for example, using "protocols" used by radio amateurs ("end", „roger", etc).



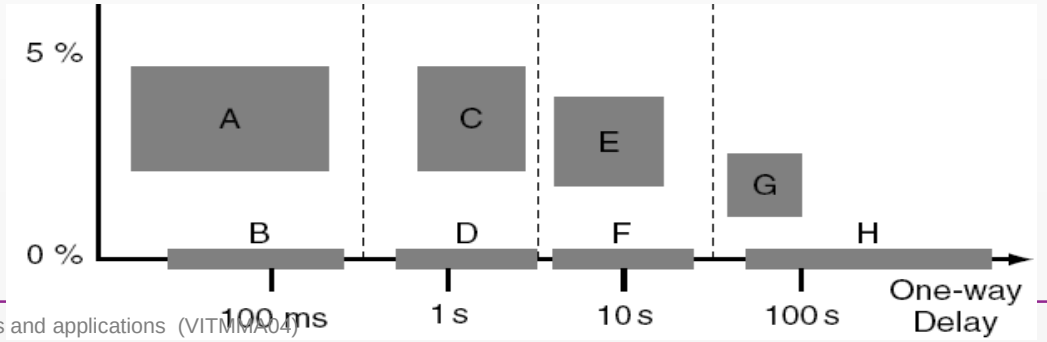
QoS target areas (cont'd)

- **Zone B: QoS for perceptual and precize media**
  - Services based on digital media, such as telnet sessions or interactive games.
  - Source content is *digital*. Loss is not tolerated, so the target is 0% for packet loss.
  - The delay typically has to stay **below 200 msec**.
  - In the case of a delay exceeding the limit, the usability of the service decreases as the user is disrupted in the conversation with the other party.
  - There is nothing to compensate for too much delay.



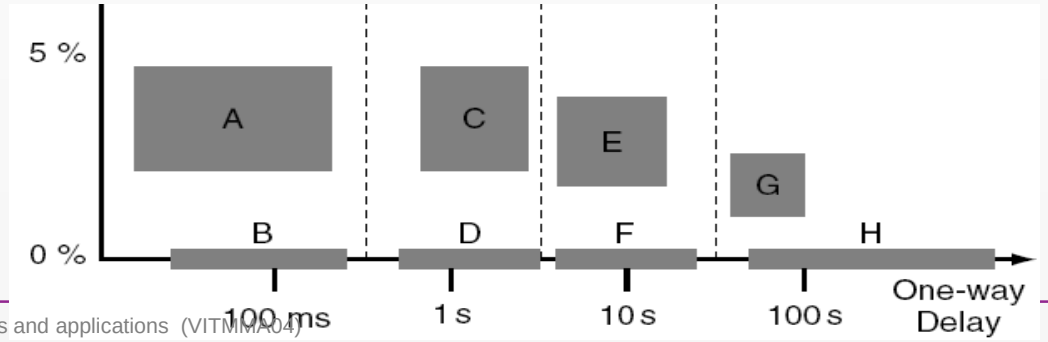
QoS target areas (cont'd)

- **Zone C: QoS for cognitive and forgiving media**
  - One-way analog services such as voicemail.
  - Similar to Zone A, but – due to being unidirectional – source content can be delayed without being detected by the user.
  - The delay is noticeable only when the stream is started.
  - Delays in **seconds**.
  - If the delay is bigger than expected, the user can quickly be notified of the effect, ensuring that the content is coming soon (eg, a message or beep that playback will start soon).



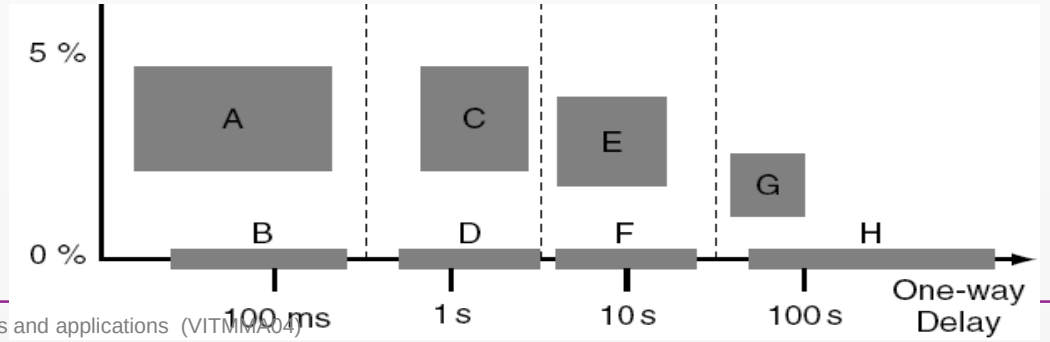
QoS target areas (cont'd)

- **Zone D: QoS for cognitive and precize media**
  - *Interactive digital services such as Internet browsing or e-business Web applications*.
  - **Loss can not be tolerated**.
  - As with computer interfaces in general, the delay should remain in the order of a **few seconds**.
  - If the response is delayed too much, short-term memory and attention wanes. For the user, the task to be performed will be more difficult (eg, you make more mistakes, your satisfaction decreases).
  - The effect of excessive delays can be reduced by feedback to the user that his/her service is being processed. Feedback can reduce frustration, but it can not extend short-term memory and can not compensate for breaks in the completed process.



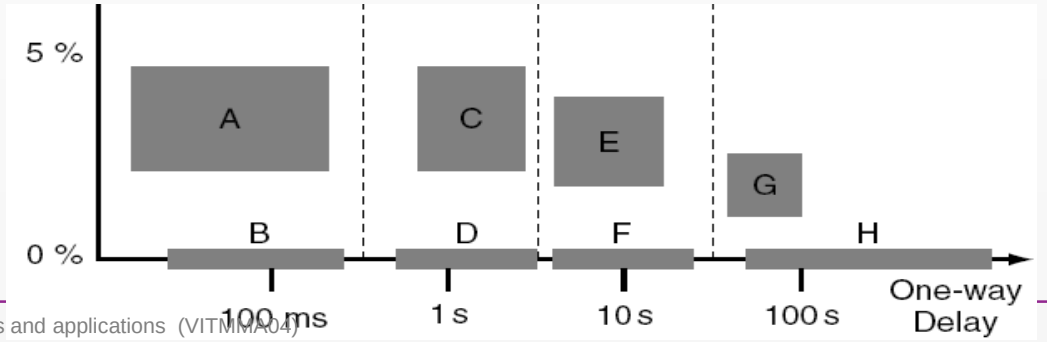
QoS target areas (cont'd)

- **Zone E: QoS for social and forgiving media**
  - Unidirectional „streaming" analog source content delivery services, like audio and video.
  - Compared to Zone C, the difference is that the content is much heavier due to its size or continuity (eg, Internet radio), so it is more problematic to restart and replay.
  - **The start-up delay may also be of 10 seconds**, as the expected time of the process is longer by a few orders of magnitude.
  - The annoyance of the user is not a neurological one, but a function of experience and expectations.
  - Feedback is very important to comfort the user at high delays (Eg, „buffering… 50%").
  - If the start-up delay is an order of magnitude higher than expected, users will assume that the service is unavailable.



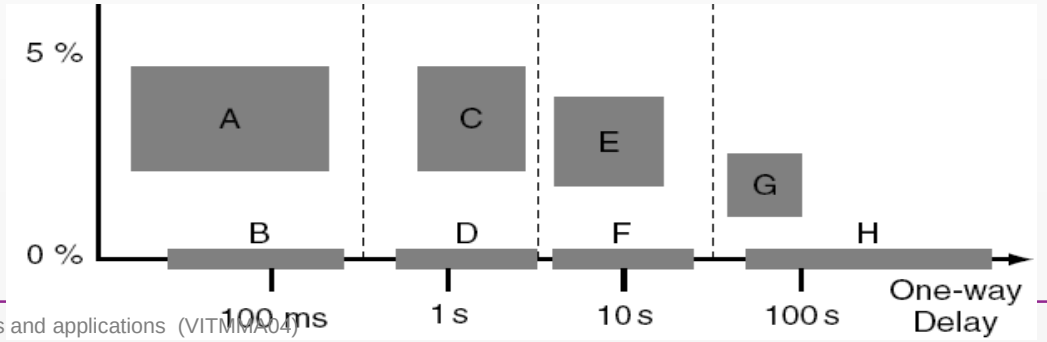
QoS target areas (cont'd)

- **Zone F: QoS for social and precize media**
  - It is similar to zone E, except that the source is digital or static (non-streaming and persistent) by nature, like image download or FTP download (eg, software downloads).
  - Contrary to zone E, a **loss is not acceptable**.
  - The start-up delay is similar to Zone E, in the order of **10 seconds**.
  - Excessive start-up delay should be treated similarly to zone E.
  - Transferring content in zone F is a finite task. Transmission is completed when the total data is received.
  - Prediction of performance allows you to display the status of the process and the remaining time for the user (eg, countdown or status bar).



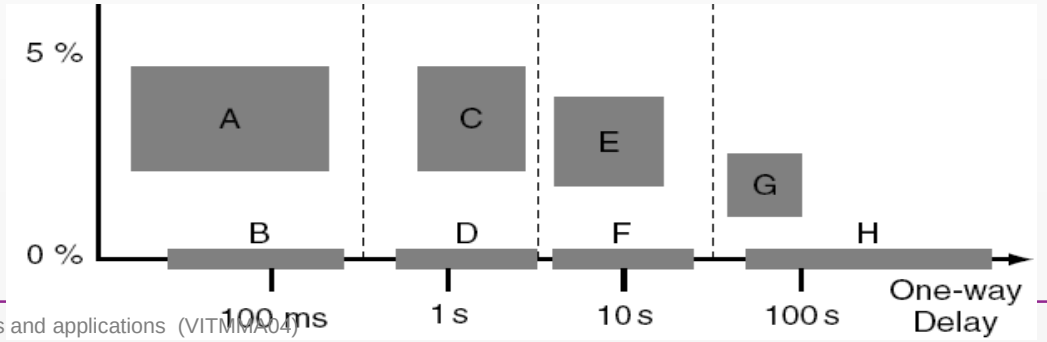
QoS target areas (cont'd)

- **Zone G: QoS for postal and forgivingmedia**
  - *Non-digital* content (such as fax) is the typical service.
  - The final product is static and persistent. For this reason, errors are more noticeable, so the **acceptable error rate is lower**.
  - The acceptable delay is much greater (somewhere around **20 seconds** and a **half minute**).
  - Unless the sender contacts the receiver immediately, the delay can not be experienced.
  - Feedback eliminates user dissatisfaction with delays that are far above the limit (eg, "the dialed number is busy, redial in 120 seconds ...").



QoS target areas (cont'd)

- **Zone H: QoS for postal and precize media**
  - *Digital services* such as e-mail.
  - **The loss must be 0%**.
  - Acceptable delay may vary widely, **from minutes to hours**.
  - Avoiding longer than expected transfer delays is typically not possible online. Allowing the user to query the status of the service can be a solution (Eg, „has the sms left?").



QoS handling in abnormal situations

Content

- SLA – Service Level Agreement
- Quality of Service (QoS)
  - QoS attributes, human factors, QoS zones
  - **QoS handling in abnormal situations**

Manage QoS in abnormal situations

- All QoS requirements must be met in "normal operating mode”
- …BUT it is also an important issue how QoS can be handled in abnormal or unexpected circumstances (eg overload, malfunction).
- **QoS in overload**
- Telecommunication systems are life-saving devices in emergencies or natural disasters.
- … BUT, the more serious or wider the emergency, the heavier telecommunications networks may be overwhelmed.
- There may be other causes of overload
  - „ mass call event”,
  - DOS attacks, etc.

QoS in overload

- Overload can easily be orders of magnitude larger than the normal operating load.
- In the case of overload, there are two basic approaches to assuring QoS:
  - **Increase system capacity**, and/or
  - **application of load control**
    - load balancing,
    - access control: e.g., load "thinning".
- Adding extra capacity can be a simple and workable strategy as long as this is technically feasible!
- The biggest threat: In the absence of load control, **QoS may crash** if the offered traffic exceeds the system's capacity!

QoS in overload

- Principles for providing QoS in overload :
  - Wherever possible, ensure a system capacity exceeding the requirements;
  - the QoS requirements must always be ensured for the traffic transmitted;
  - the carried traffic shall be treated equally without distinction;
  - in the case of traffic exceeding capacity, the traffic shall be separated: the part to be transferred and not to be transmitted;
  - segregation of traffic (eg, access control, load shedding) should be done as close as possible to the source.

QoS in malfunctions

- Telecommunication networks are designed to maximize reliability for potential hardware and software malfunctions.
- If a service can be restored after the failure, the question is *how quickly* we can do this
  - For example, redirecting or involving reserve resources?
- This restoration time will cause additional delays for the service.
- Typically, for any streaming audio or video in "perception category", this delay will lead to interruptions that the user will detect...