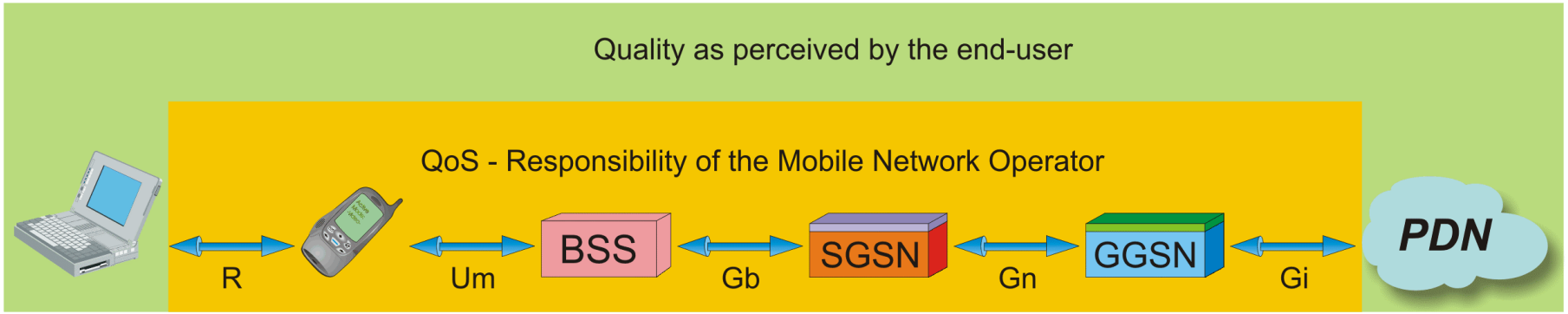
What is Quality of Service in GPRS Networks

- **The QoS-Scope of the GPRS Network Operator**



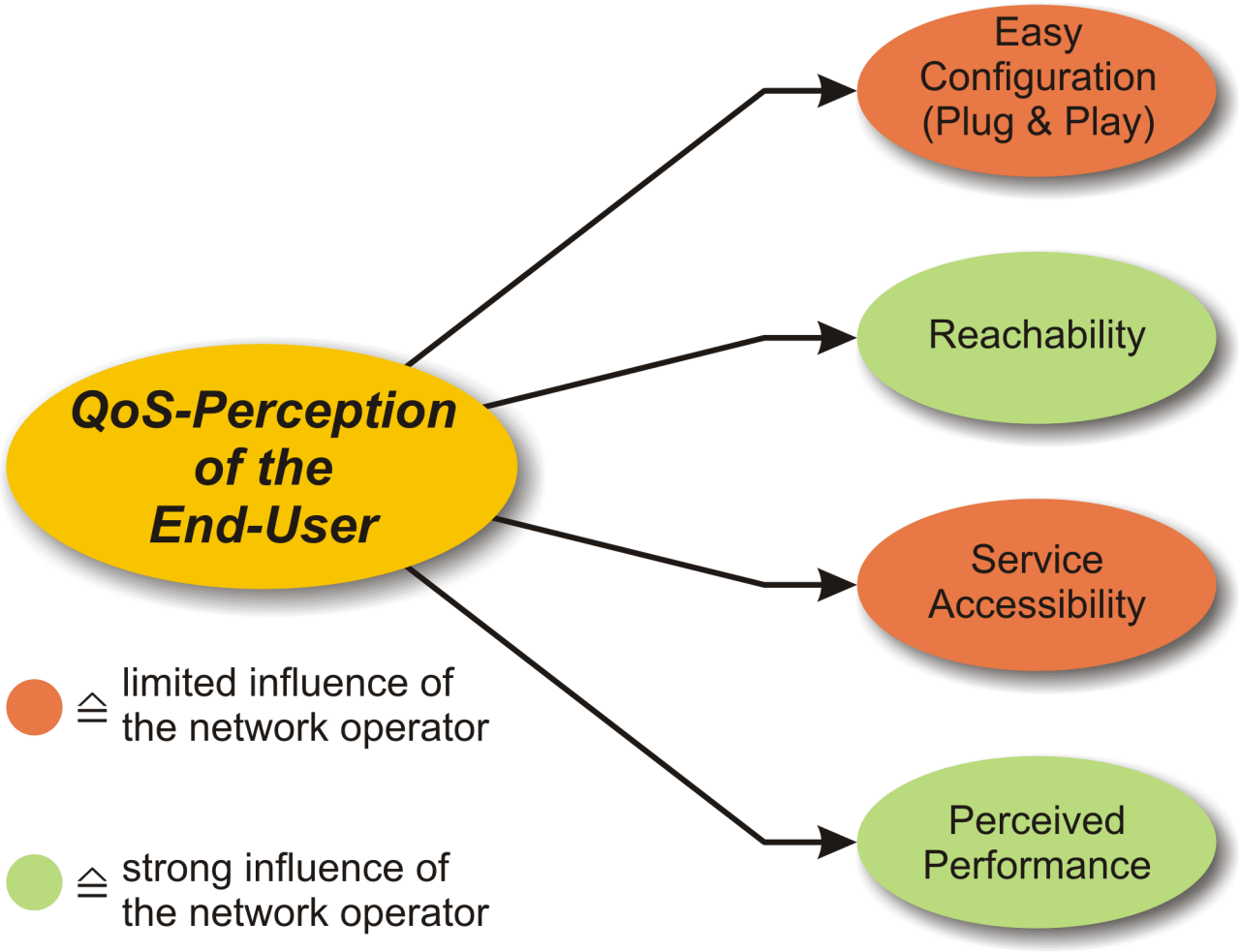
What is Quality of Service in GPRS Networks

The QoS-Scope of the GPRS Network Operator

GPRS network operators are obviously eager to provide the best service possible to their customers. However, the figure already emphasizes the problem: While the end-user has an end-to-end view of performance and quality, the GPRS network operator can only be responsible for his part of the data transmission chain.
For example, delays which are caused by congestion on the internet or badly configured terminal equipment will limit the performance essentially and the customer will definitely blame it on the GPRS network.

It is particular this perception of the end-user which forces the GPRS network operator to provide an optimum performance and to continuously measure this performance by suitable measures.

The QoS-Perception of the End-User



The QoS-Perception of the End-User

Easy Configuration

Nowadays, one of the most important hurdles to use GPRS is the complicated setup of GPRS in the terminal equipment, most likely a laptop and in the mobile station itself. Even engineers have to struggle with the task of configuring GPRS. Apparently, nobody really considered AOL's advertisement slogan properly: "Oh, I'm already in …" (the internet). In the long term, this needs to change.

Reachability

The most important feature of GPRS is the "Always On" feature. At least in theory, GPRS allows the customer to stay on the internet continuously. However, reality is different. Another issue is the lacking possibility to push IP-content to mobile subscribers. This would require the mobile terminating way of PDP-context activation.
Likewise, mobile subscribers need to stay reachable for their circuit-switched partners (voice calls) even while surfing the web. This can only be achieved through the introduction of NOM I.

Service Accessibility

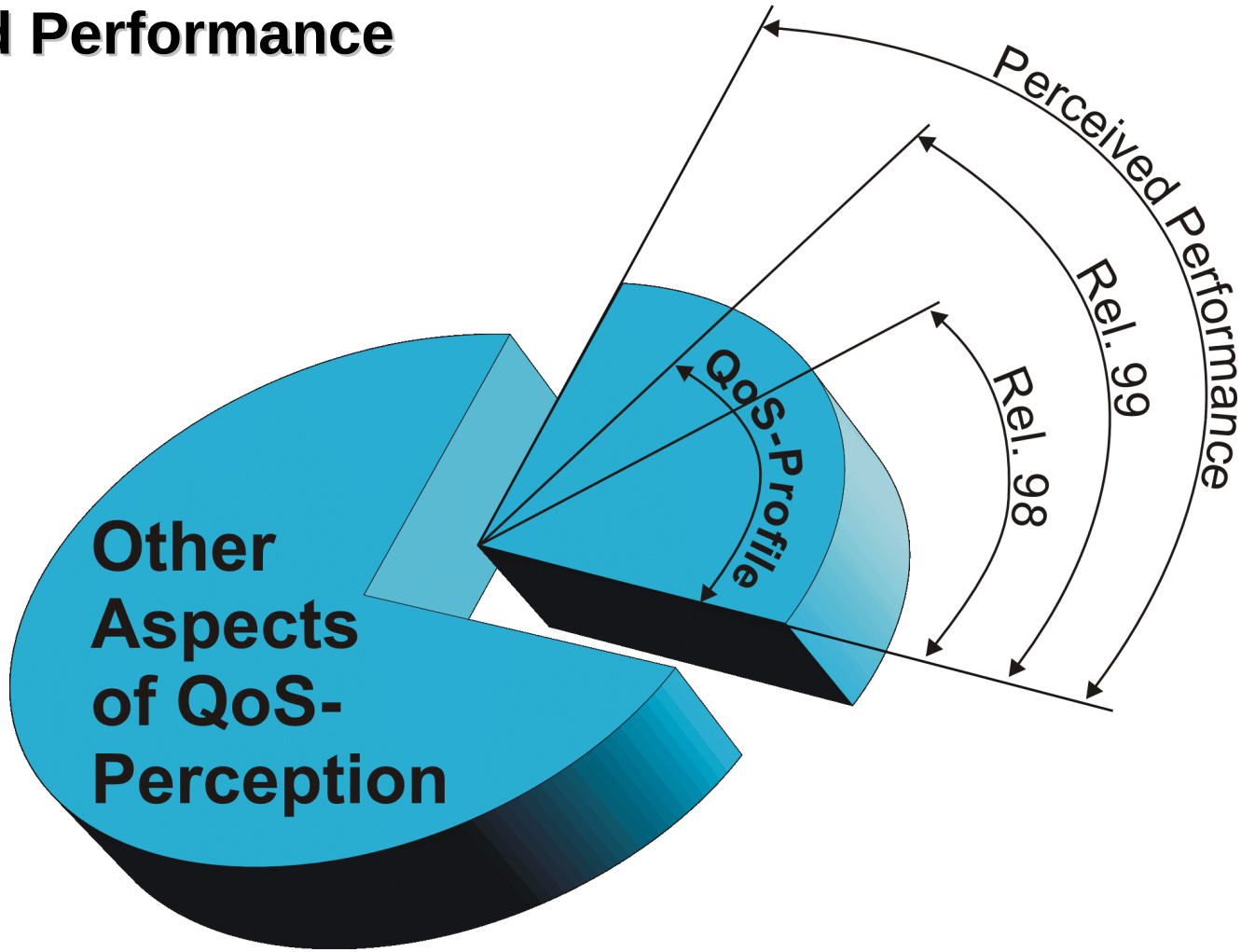
Today, GPRS is primarily the domain of business users who travel abroad. Unfortunately, GPRS does still not stable operate while roaming outside the Home-PLMN. Adjusting GPRS-settings while roaming cannot be the final solution.
Some applications are basically out of scope for GPRS users. Surfing the internet with PDA-devices is no real pleasure because of the limited user interface (screen, keyboard) and for some applications GPRS is simply to slow.

Perceived Performance

For the end user the performance is basically how much content can be transmitted in what time. GPRS users will also soon recognize the inherent latency of GPRS to translate a user input into a reaction, into a data transmission.

The following chapters will primarily deal with the measurement, consideration and optimization of this perceived performance.

The Relationship between QoS-Profile, QoS-Perception and Perceived Performance



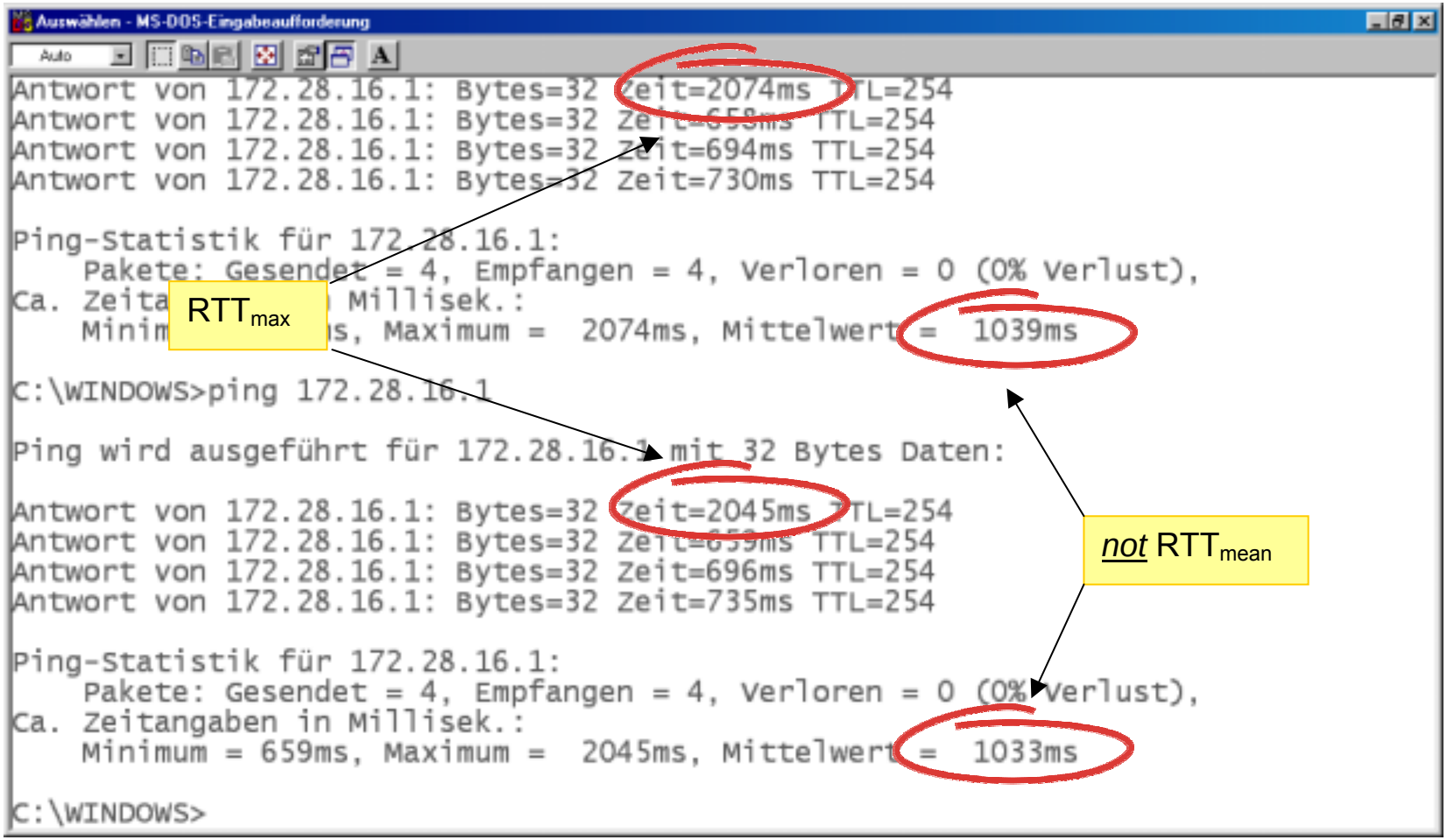
The Relationship between QoS-Profile, QoS-Perception and Perceived Performance

The diagram illustrates the relationship quite nicely:

⇒ Neither the QoS-profile Rel. 98 nor the QoS-profile Rel. 99 do match the perceived performance (⇔ how much content in which time).

⇒ And the QoS-perception of the end-user by far exceeds any definable QoS-profile.

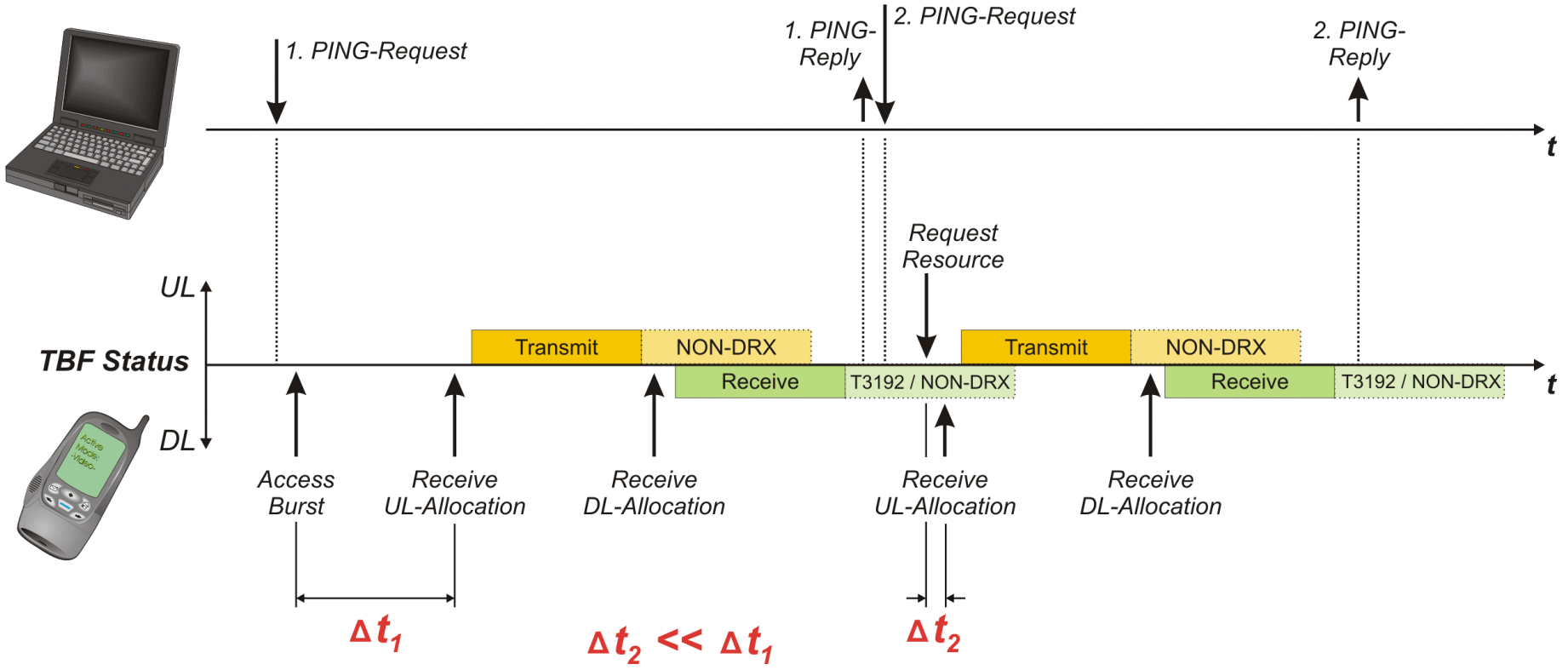
Maximum and Mean Round Trip Time



Maximum and Mean Round Trip Time

The screenshot illustrates the measurement of RTTmean and RTTmax using the ICMP PING-application. Please note the essential discrepancy between the initial response from the router and the following responses.

Reasons for the Huge Difference between RTT_{mean} and RTT_{max}

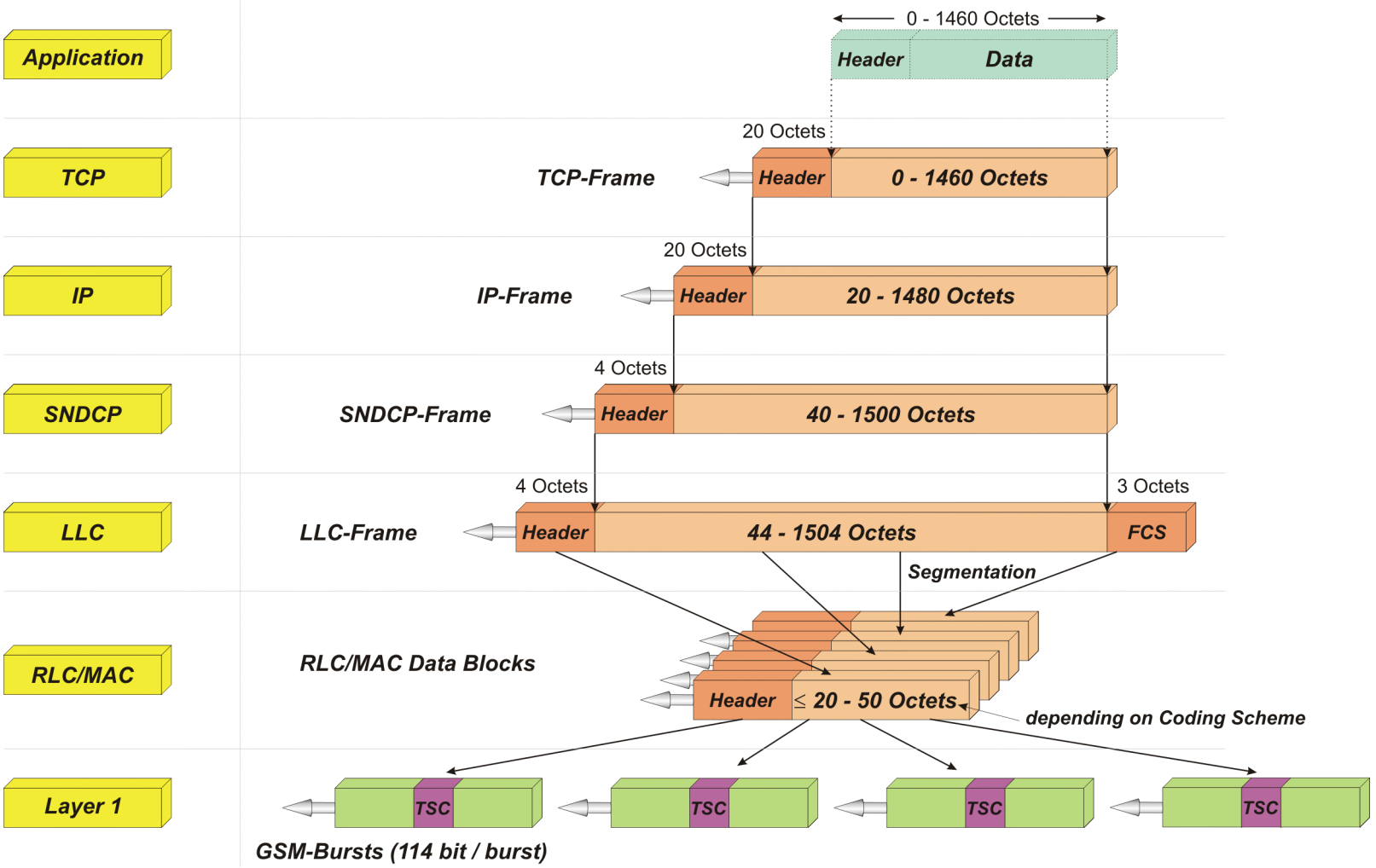


Reasons for the Huge Difference between RTT_{mean} and RTT_{max}

Before we continue let us investigate in more detail what causes the huge difference between RTT_{mean} and RTT_{max}:

- ⇒ We assume that the mobile station is currently in packet idle mode.
- ⇒ Initially, the mobile station will receive the PING-Request from the terminal equipment.
- ⇒ To suit this transmission request, the mobile station will send an access burst to the network, asking for the allocation of uplink resources.
- ⇒ As can be seen, the time Δt_1 is quite long and varies with the used access method (One-Phase / Two-Phase), with the network load and it depends on whether the access is performed using PRACH or RACH.
- ⇒ The mobile station will use the allocated uplink resource to convey the PING-Request to the PCU which will forward it through the network. Finally, the PING-Request reaches its destination (the router behind the GGSN in this case).
- ⇒ This router will respond to the ECHO-Request by sending an ECHO-Reply to the terminal equipment.
- ⇒ When this ECHO-Reply reaches finally the PCU, the uplink resources are obviously already released but the mobile station is still in NON-DRX-mode and in GMM-Ready State (depending on the setting of the respective timers).
- ⇒ Accordingly, the mobile station does not need to be paged but receives a direct resource allocation which speeds up the process. On the allocated DL-resources, the PCU will convey the ECHO-Reply to the mobile station which in turn will forward it to the terminal equipment.
- ⇒ Having received a reply for its initial request, the terminal equipment will send the next ECHO-Request.
- ⇒ Depending on whether the downlink TBF is still active or whether T3192 or the NON-DRX-timer are currently running, the mobile station will ask for uplink resources through another access procedure or by using the uplink PACCH of the downlink TBF.
- ⇒ Because of this, the mobile station may obtain the resources faster and hence Δt_2 is much smaller than Δt_1.

Performance Leakage within the GPRS Protocol Stack



Performance Leakage within the GPRS Protocol Stack

The figure illustrates the performance leakage of the GPRS protocol stack and the performance leakage which is caused by the TCP/IP-headers.

RLCMAC
On RLC/MAC level, the maximum throughput rate depends on the available coding schemes (CS-1: 8 kbit/s x TS; CS-2: 12 kbit/s x TS; CS-3: 14.4 kbit/s x TS; CS-4: 20 kbit/s x TS). Please refer to [1] for more details.

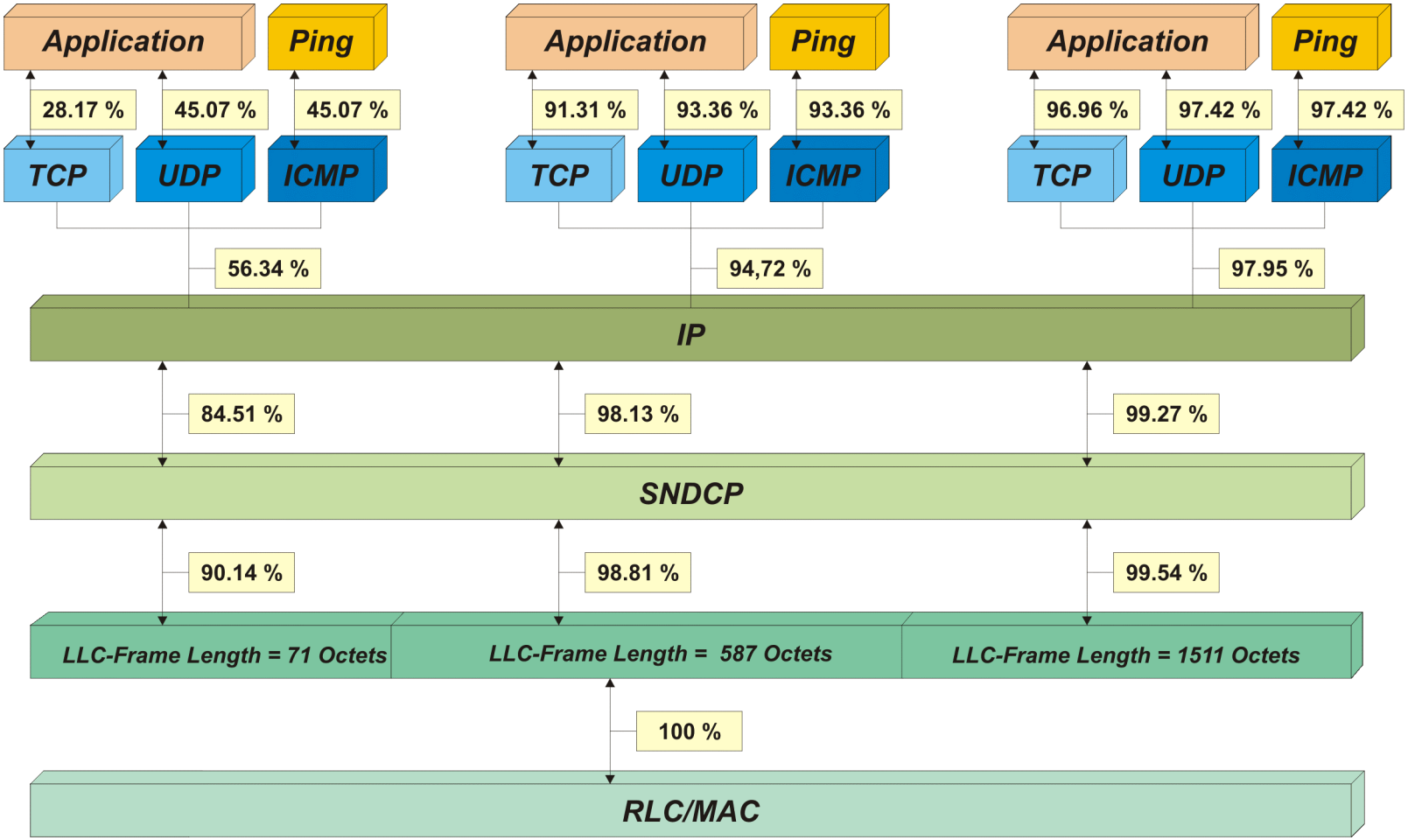
LLC
On LLC level, the maximum throughput rate depends on the maximum size of the information field which is controlled by the parameter N201 and on the type of LLC-frame which is used for data transmission. Throughout the class we will assume that LLC-UI-frames are used (unacknowledged transmission).

SNDCP
SNDCP will add another header to the IP-frame.

IP
For the internet protocol, we need to consider another header with 20 octets length. Optional fields of the header may possibly need to be added.

TCP
For the transmission control protocol, we need to consider another header with 20 octets length. Optional fields of the header may possibly need to be added.

Relative Performance Leakage

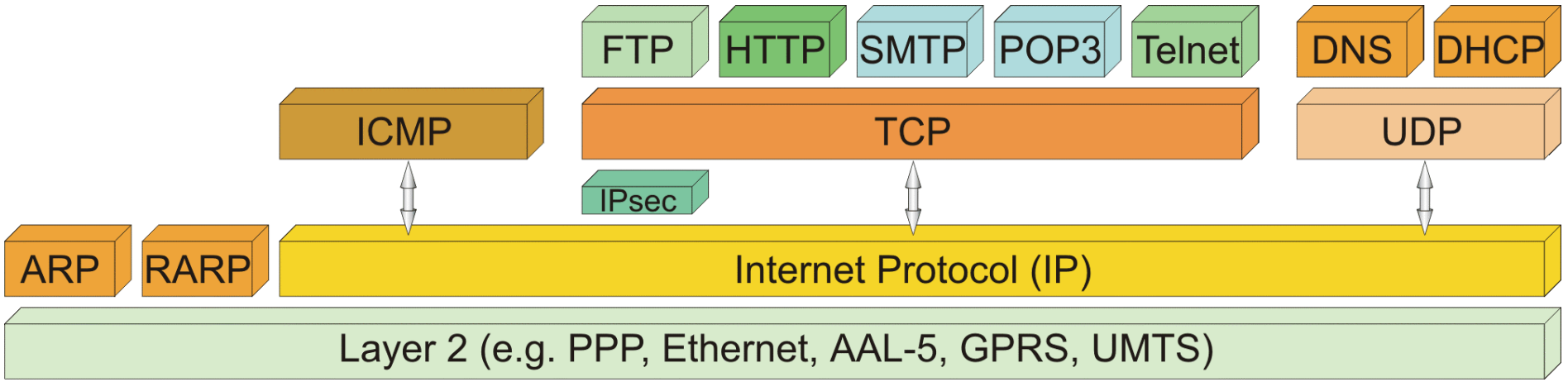


Relative Performance Leakage

The relative performance leakage considers the header lengths from the previous pages but adds a few more details:

⇒ The figure illustrates three columns which start at the LLC-layer and which depend on the actual length of the LLC-frame.
⇒ We also added UDP and ICMP as layer 4 protocols in addition to TCP (we apologize for calling ICMP a layer 4 protocol)

Introducing the IP-Protocol Stack



Introducing the IP-Protocol Stack

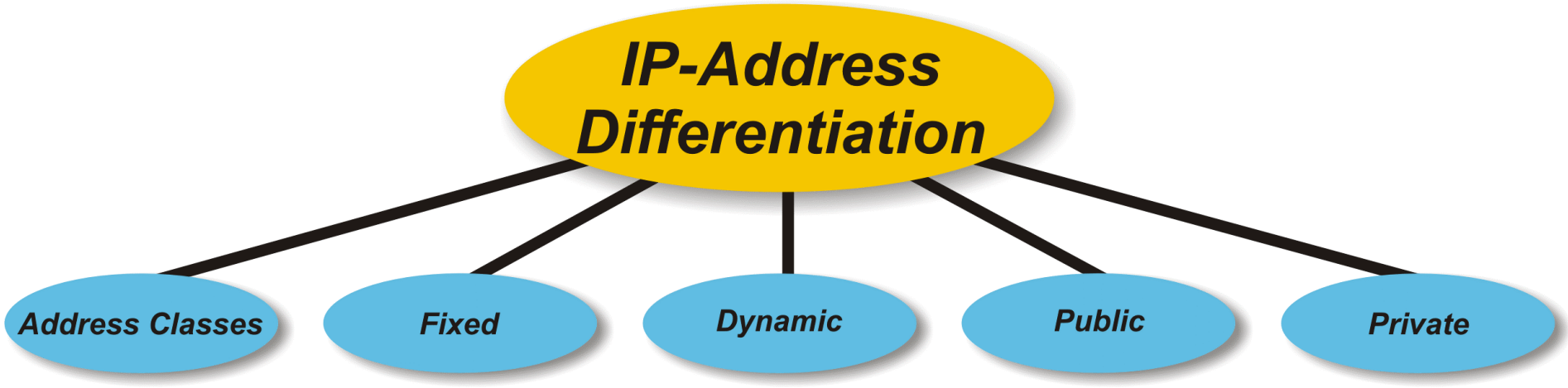
A simplified view of the IP-protocol stack is provided in the figure. The Internet Protocol (IP) itself can be located on top of almost any available layer 2-protocol like for example:

⇒ PPP (⇔ Point-to-Point Protocol / RFC 1661)
⇒ Ethernet (⇔ IEEE 802.3)
⇒ AAL-5 (⇔ ATM / ITU-T I.363.5 (6), Q.2110 (4)
⇒ or even on mobile bearers like GPRS or UMTS (⇔ 3GPP recommendations).

This flexibility of IP makes it the preferred network layer solution of the starting 21st century.

However, the term IP-protocol stack rather relates to IP and the higher layers which make use of the IP's networking capabilities. In the following sub-clauses we will take a more detailed look at the IP itself and the protocols on top of it.

IP-Addresses



IP-Addresses

One of the key issues of the internet protocol are the 32 bit long IP-addresses. The IP-address shall identify any given user of the internet uniquely at any given moment in time. Due to the enormous growth of the internet community, IP-addresses are becoming a scarce resource. Special features have been put in place to make more efficient use of the available IP-address range. IP-addresses can be distinguished by the following means:

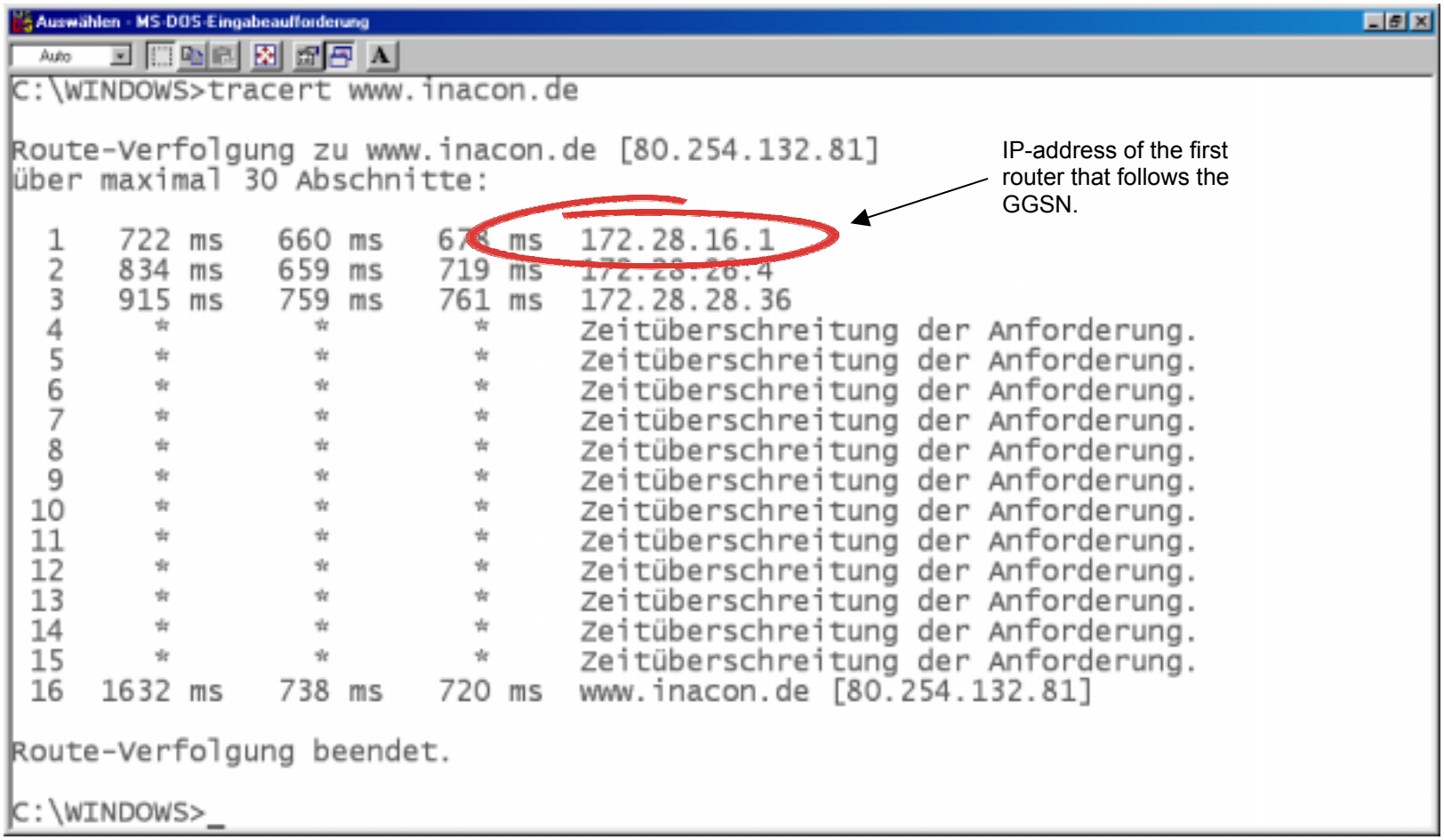
- **Address Classes**
  Already at the very beginning of the internet, five different address classes (A, B, C, D, E) were defined which imposed a hierarchical and unfortunately uneconomic structure upon all internet addressing. The ICANN (Internet Corporation for Assigned Names and Numbers) is now in charge for allocating IP-addresses to individuals and organizations. Please note that ICANN replaced IANA (Internet Assigned Number Authority) in 1999.

- **Fixed and Dynamic IP-Addresses**
  Any IP-address is either fixed or dynamic. In the “fixed” case, each user has at least one IP-address which cannot be used by anybody else even when this user has switched off the computer. Opposed to that, when dynamic addressing is used, any user of a system will obtain another IP-address any time the host computer is switched on. Switching off the computer also means to return the IP-address, used, and therefore making it available again to another user. We will later get back to the issue of dynamic IP-addressing.

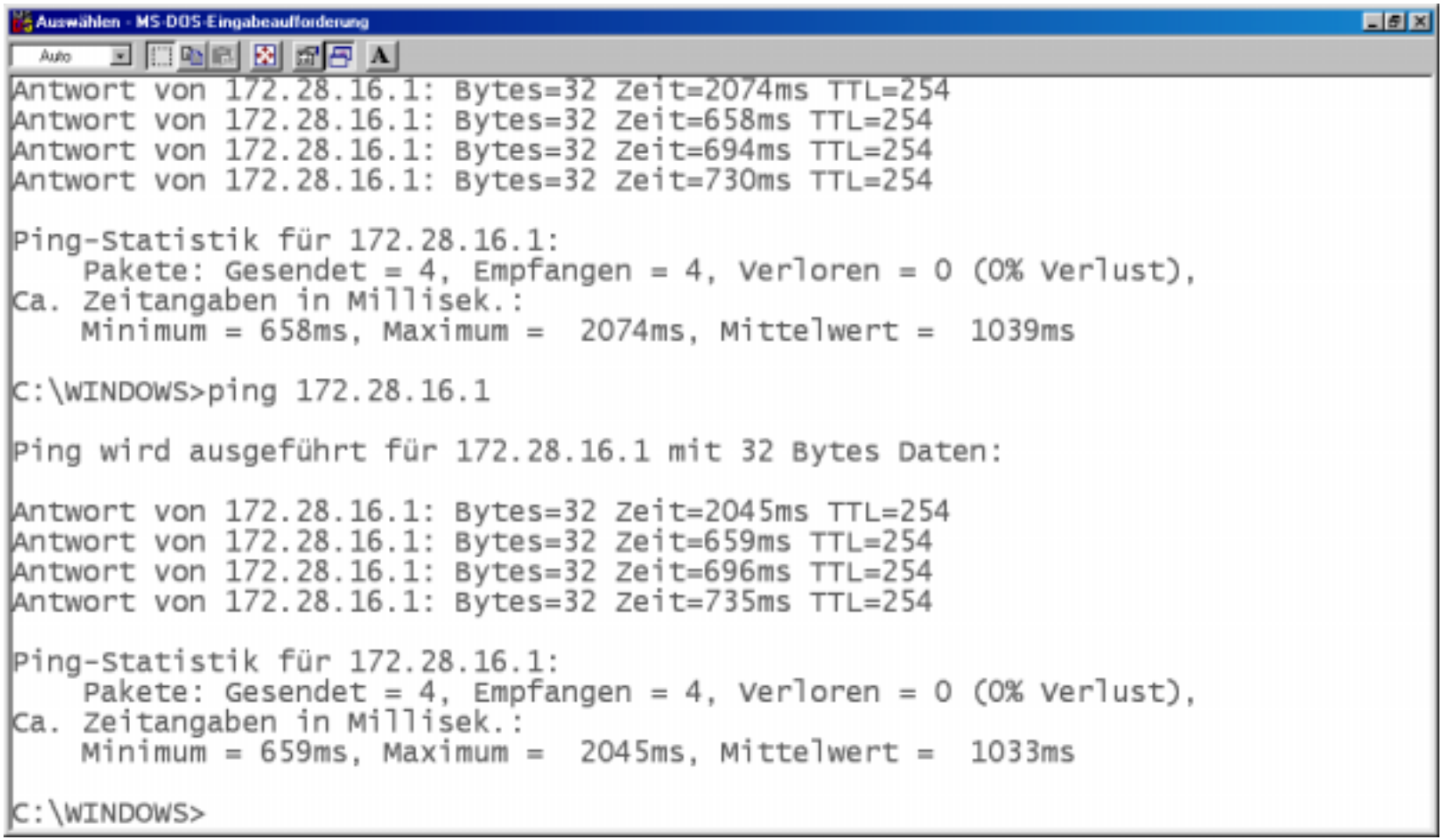
- **Public and Private IP-Addresses**
  Yet another means to distinguish IP-addresses is to have either public or private IP-addresses. Private IP-addresses are unique only and exclusively in the very network where they are used. However, such an IP-address cannot be used for communication purposes towards hosts or servers outside this network. Special means are required to allow such users access to services outside the local network.
  Opposed to that, public IP-addresses are unique worldwide and can be used for any access to services inside and outside the local network. We will later see more details about private and public IP-addressing.

In this context we will focus on the 32 bit long IPv4-addresses. In addition, the IETF and major industry leaders are preparing the roll-out of IPv6 [RFC 1883], also referred to as IPnG (⇔ for next generation). IPv6 comes with the major advantage of using IP-addresses with a length of 128 bit compared to 32 bit in IPv4.

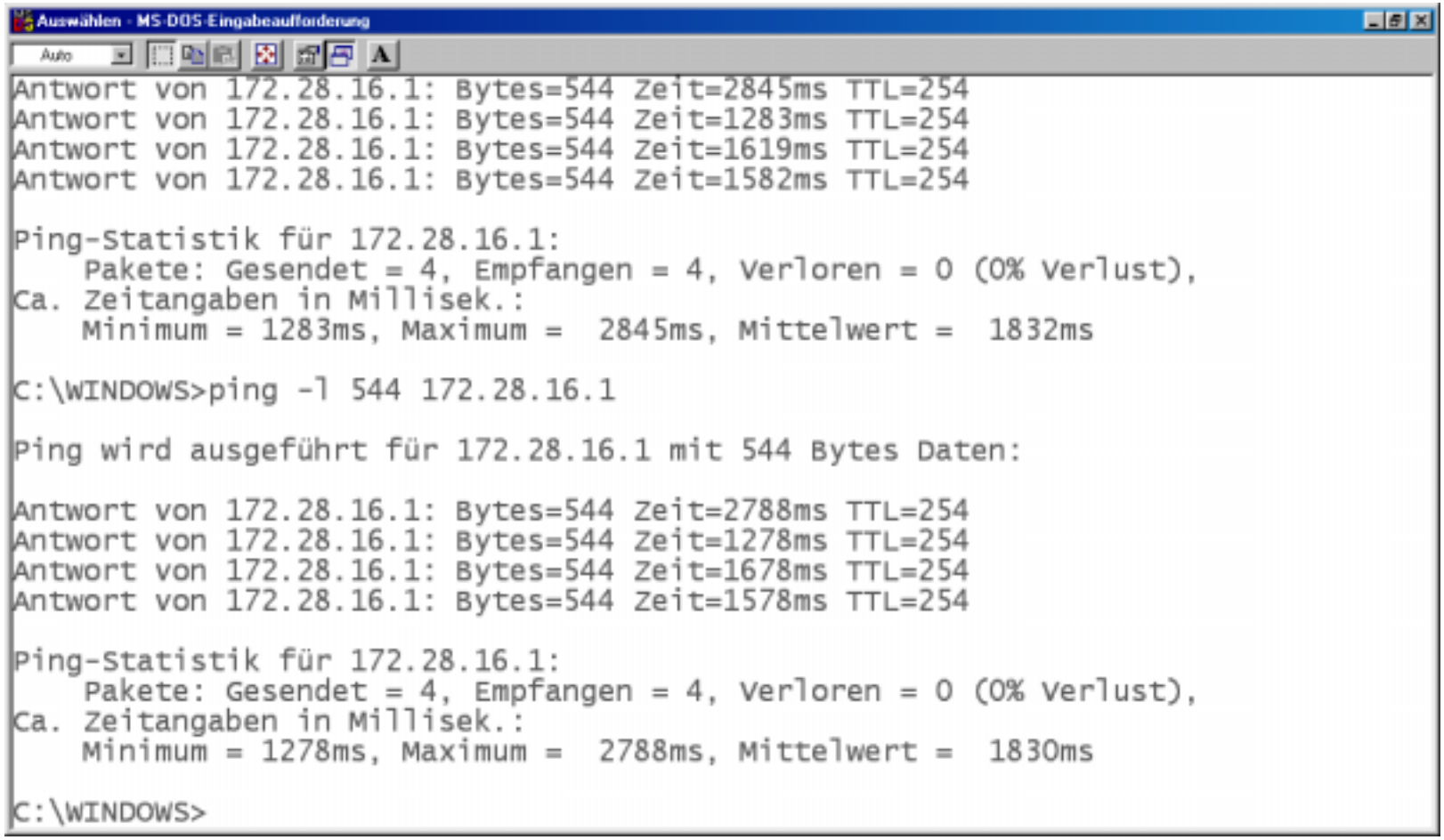
Use Trace Route to Determine the IP-Address of the 1st Router



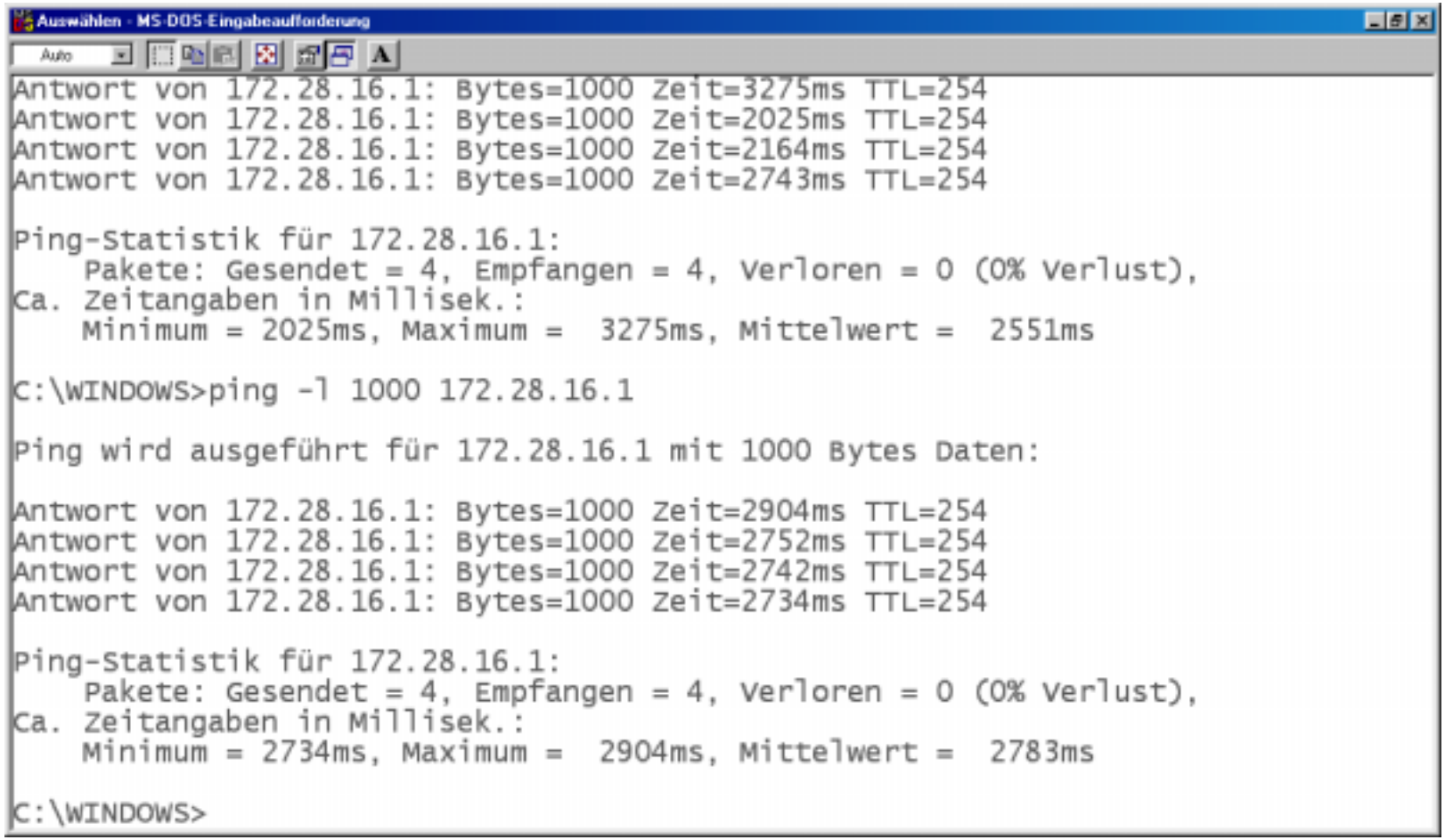
Ping with 32 Octets of Data (no Segmentation)



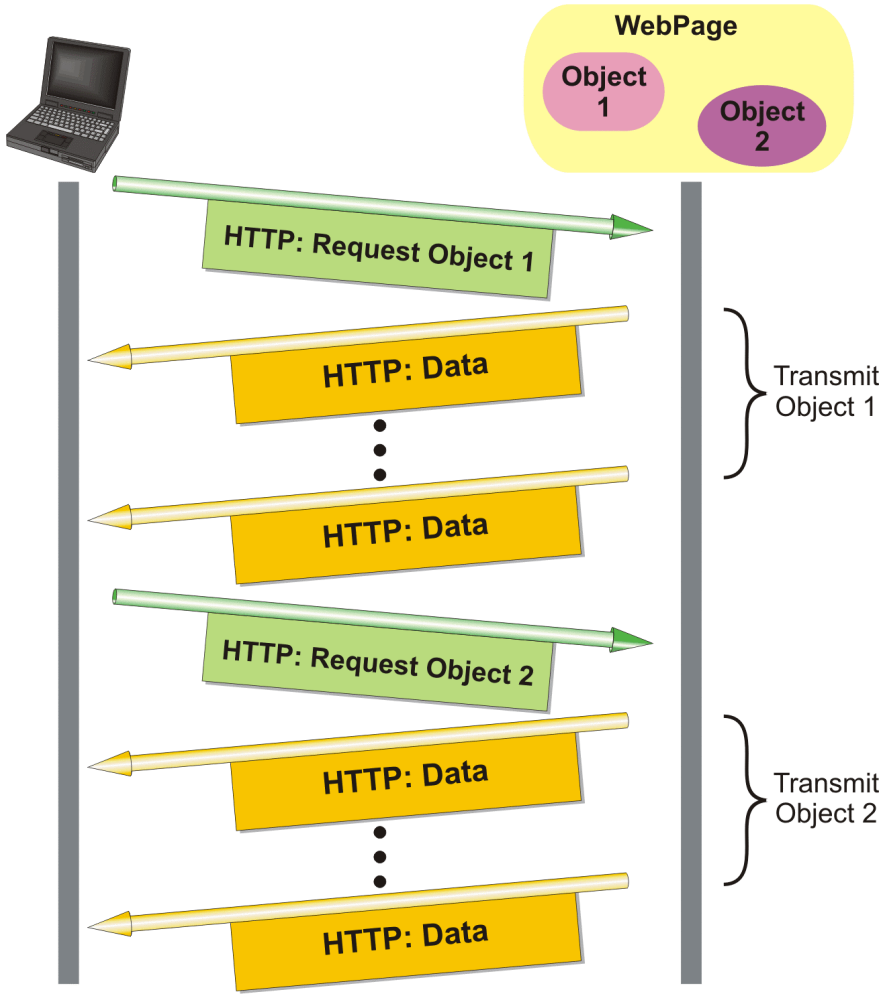
Ping with 544 Octets of Data (still no Segmentation)



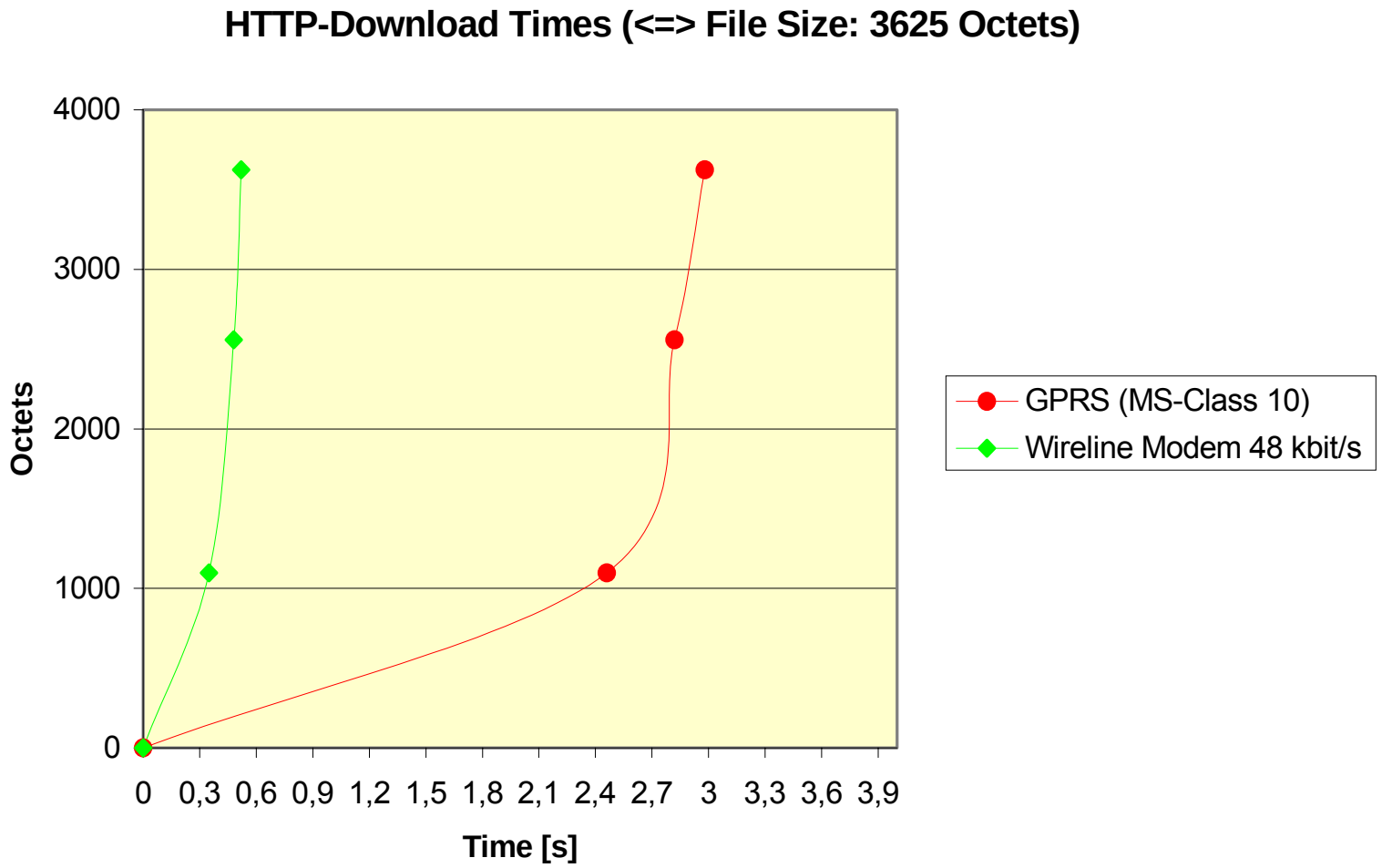
Ping with 1000 Octets of Data (Segmentation)



Download of a given Web Page



Impact of GPRS Specific Delays on HTTP-Performance



Index:

N

O

P

R

S

T

U

V

W