User Traffic Modeling for Future Mobile Systems

- The goal was to gain new knowledge and develop expertise about the fine structure functionality of packet data traffic for the development of future mobile data systems.
- To create a model, which adequately describes the characteristics of the individual user’s connection over different time scales.
- A special interest in the lower levels of the time scale
- Packet data traffic measurements for
  - WWW service in laboratory LAN 1996 and 1999 and WLAN 1997.
  - WAP over GSM data and over GPRS from a test WAP-gateway 2000-2002
- The data was grouped by individual connections and analyzed based on the protocols used.
- The statistics were modeled based on events of WWW session
  - Developed from the ETSI packet data model (referred Ch. 10.1)
  - measured distributions are fitted to some analytic distributions
  - aim is to get parameters for simulation model(s)
  - intended to developing radio link protocols and radio network planning

Wired vs. Mobile Data Traffic

- In fixed networks
  - bandwidth is large and rapidly growing and transmission errors are rare
  - most crucial elements are the centralized components like main trunks, routers or servers
  - one of the main problems is the aggregate traffic of numerous users, which overloads these relatively few "bottlenecks"
    => the traffic should be measured from the "hot spots".

- In mobile networks
  - Bandwidth is quite limited and the probability of transmission errors is rather high
  - Few active users can make use of most of the traffic capacity available in a cell
  - The main "bottleneck" is the air interface at the edge of network
    => the traffic should be measured as close to the client as possible.

- In WCDMA BER/FER performance is optimized based on average E_b/N_0
  - The average E_b/N_0 is not accurate if high bit rate packet users cause rapid changes in interference.

WWW traffic

- One of the most spread services in the Internet
- Often used as user interface for new services
- HTTP protocol
- Uses TCP and IP protocols for transmission

- The technology develops on various levels => has impact on the results
  - Internet bandwidth is increasing
  - Processing power of both clients and servers is increasing
  - New software versions offer more capabilities

- Changes in the user behavior and the contents of Internet
  - Amount of data in Internet is increasing
  - People use WEB more frequently
  - Number of items per page is increasing

- Physical distances remain => Round trip time (RTT)

The UMTS-network

- aimed to cover almost all the data transmission needs of the users
- different delay and other quality demands
- the behavior of most significant services present in the network is needed to
  - follow the effects of changes loading
  - evaluate the functionality of the network
  - evaluate the service quality (see lect. 1 p. 32-36)
  - control them (for example the usage of priorities)

WAP traffic

- to provide a mobile user a WWW like access to the Internet.
- a HTTP-like protocol optimized to the wireless domain.
- Uses TCP and IP protocols for transmission

- The measurements used circuit switched GSM data and WAP protocol 1.0.
- The traffic logged simultaneously from both sides of the gateway.
- The effects of wireless and Internet connection and the gateway separated
- already the activity during WAP-transaction < voice activity (esp. uplink)

=> advantages by multiplexing several sources to shared channels.

- implementation errors generate "pseudo traffic"

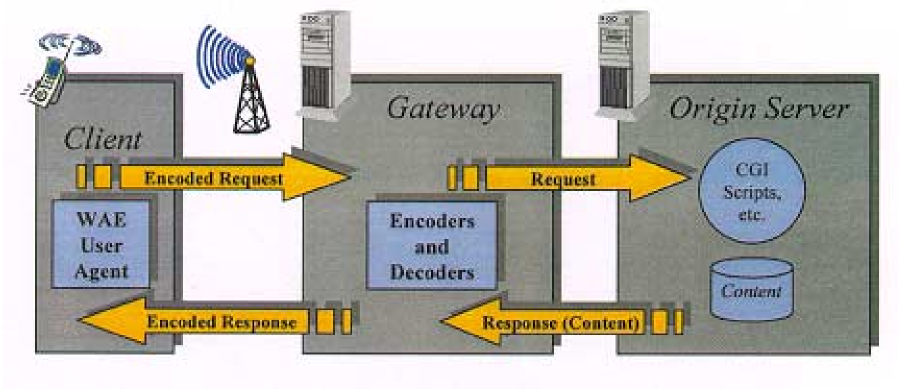

Figure 1: The WAP Environment

- The normal structure for a WAP-transaction is
  1) WAP-request,
  2) WWW-connection
  3) WAP-response
- Delay is always larger than with WWW-connection in fixed network

- Two protocols used
  - Wireless Transaction Protocol (WSP/WTP UDP) port number 9201 and
  - Wireless Datagram Protocol WSP/WDP UDP port number 9200.

Packet data traffic measurements

- Data packets were collected by TCPDUMP
- analyzed by C and MATLAB programs.
- The data was grouped and analyzed based by
  - Users (or PC) indicated by IP- and MAC-addresses
  - Services indicated by ports used by TCP/UDP-protocols.

- Packet level statistics 42 figures
  - size of every IP- and data packet both directions
  - delays between packets in both directions
  - comparisons of delay distributions
  - delays between packets on the same WAP-item
  - number of WWW-items/page and -pages/session
- Bursts, Nibbles, WAP-items, -connections, -pages and WAP-sessions, 24 figures each
  - size of groups in packets and in bytes
  - delay from previous group
  - length of the group
  - cumulative distributions
  - distributions of bytes based on length of group

The used definitions

Packet

IP-packet

Nibble

Smallest burst of data, which UMTS would distinguish.
group separated with idle > 10 ms.

Burst

active transmission, group separated with idle > 2 s.

TCP/ WSP-connection

A numbered connection/transaction between WWW-server and -client or WAP-gateway and -client.
One TCP-connection can carry tens of WWW-items.

WWW/WAP-item

A request/response pair transferring NEW payload data text, picture etc., on same TCP/WSP-connection

WWW/WAP-page

WAP-items that forms one visual display unit. Separated by a reading period, defined from 1 to 300 seconds.

WWW/WAP-session

A period when client is active. Separated by inactivity of no WWW/WAP-page during 5 minutes (= reading time > 5 min.)

Creating the model

- Modeling is done by fitting the cdf of the result to analytic distributions / their mixtures.
- To maintain the information over different time scales, the fitting is done using logarithmic x-axis
- A discrete vector of size 221 samples covers the time scale from 10^{-6} (1 μs) to 10^{5} (1,25 days) with a resolution of 20 points/decade.

- The model is fitted to the measured distribution by numerical iterations.
- The correctness of fitting is evaluated visually
- The distributions used are exponential, Pareto and for small discrete values also geometric.

- the distributions have been enhanced to fit better to the measured data
- no zero length delays => shift (= fixed delay) added to exponential distribution
- bias (= fixed value of zero)

The Pareto distribution

- is defined by

$$f_x(x) = \frac{\alpha \cdot k^{\alpha}}{x^{\alpha+1}} \qquad , x \geq k \qquad (0.1)$$

$$F_x(x) = \begin{cases} 0 & , k < x \\ 1 - (k/x)^{\alpha} & , k \leq x \end{cases} \qquad (0.2)$$

$$\mu = \frac{k\alpha}{\alpha - 1} \qquad , \alpha \geq 1 \qquad (0.3)$$

$$\sigma^2 = \frac{k^2 \cdot \alpha}{(\alpha-2)\cdot(\alpha-1)} \qquad , \alpha > 2 \qquad (0.4)$$

- when α < 2 the variance and when α < 1 also the mean become infinite
- normally the Pareto distribution is limited to area 1 < α < 2

Truncating the Pareto distribution

- parameter T added to compress the in principle unlimited Pareto distribution to the practice

$$F_x(x) = \begin{cases} 0 & ,k < x \\ \dfrac{1-(k/x)^{\alpha}}{1-(k/T)^{\alpha}} & ,k \le x \le T \\ 1 & ,x > T \end{cases} \qquad (0.5)$$

- closes unlimited Pareto, when T/k and α increase
- if k = 10^{-3} , the difference in cdf between T = 10^{3} and T = 10^{333} (~ infinity) is
  - only 10^{-9}, when α = 1.5
  - but 10^{-3}, when α = 0.5
- in many cases small values of α (min = 10^{-5}) give a pretty good fit to measured data. Then the graph becomes a slope line in semi logarithmic domain.

Geometric CDF

- directly from Matlab defined as

$$F(x \mid p) = \sum_{i=0}^{floor(x)} pq^i \qquad where \;\; q = 1 - p \qquad (0.6)$$

- Since the mathematical distribution starts from zero to reach the aimed mean P must be set

$$P = \frac{1}{1 + mean} \qquad (0.7)$$

The developed traffic data models

- The selected statistics were fitted to analytic distributions
- simple model is one CDF and partial model is weighted sum of one exponential and two truncated Pareto CDFs
- models are a collection of several measurable distributions on different levels in top-down order
- the mean and variance for the measured data and the models
- error value used as the measure in curve fitting

WWW-traffic data model

- model is a collection of eleven measurable distributions on three levels as described in figure 3 in top-down order:

1. The WWW-session interarrival time D_{www}
2. The number of packet calls (pages) per WWW-session N_{pc}
3. The reading time between packet calls (WWW-pages) D_{pc}
4. The number of items per WWW-page N_{i}.
5. The time intervals between items belonging the same WWW-page D_{pii}

6. The number and size of packets belonging to an WWW-item are conducted about the information about the TCP-protocols mechanisms and their influences and the distributions of

- ***6.1.*** WWW-item sizes on Uplink S_{iu}
- ***6.2.*** WWW-item sizes on Downlink S_{id}

7. The time intervals between packets belonging the same WWW-item are divided in four subcategories to adapt to the different delay behavior depending on the direction of transmission

- ***7.1.*** the time int. between two consecutive Uplink packets inside an item D_{iuu}
- ***7.2.*** the time interval from Uplink to Downlink packet inside an item D_{iud}
- ***7.3.*** the time int. between two cons. Downlink packets inside an item D_{idd}
- ***7.4.*** the time interval from Downlink to Uplink packet inside an item D_{idu}

To make comparison easier each distribution for both models are presented in a table for both 1996 and 1999 measurements. In a third table there is a comparison of the mean and variance for the measured data and the both models. There is also the error value, which was used as the measure in numerical curve fitting and optimization. It is the sum of squared error between the CDF vectors for measured data and the model.

A sample distribution for D_{www}

4.1 The WWW-session interarrival time, D_{www}

D_{www}	1996		1999		
Distribution	%	Parameters	%	Parameters	
Truncated k=	100	346.7	100	260.52	s
Pareto α=		0.3675		0.4195	
T=		3.32e+06		3.8236e+33	s

Table 4.1 The simple model for WWW-session interarrival time **D_{WWW}**

D_{www}	1996		1999		
Distribution	%	Parameters	%	Parameters	
Exponential μ=	3.18	3136.6	16.14	254.90	s
start is shifted		0.001		301.51	s
Truncated k=	70.46	295.9	74.53	257.93	s
Pareto α=		0.3842		0.3766	
T=		1.012e+20		1.155e+19	s
Truncated k=	26.36	643.51	9.33	1498.9	s
Pareto α=		0.4624		0.2129	
T=		7.103e+16		73509	s

Table 4.2 The partial model for WWW-session interarrival time **D_{WWW}**

D_{www}	1996			1999		
Distribution	Measured	Simple	Accurate	Measured	Simple	Accurate
Mean	16877.9	16639.4	16791.9	13358.3	14005.4	14005.0
Variance	31053.0	30805.4	31346.9	26959.0	29103.0	28691.1
Error		0.010323	0.004739		0.011458	0.004435

Table 4.3 The mean, variance and modeling error for WWW-session interarrival time **D_{WWW}**

WAP-traffic data model

- model is a collection of twelve measurable distributions on three levels as described in figure 3 in top-down order:

1. The WAP-session interarrival time D_{wap}
2. The number of packet calls (pages) per WWW-session N_{pc}
3. The reading time between packet calls (WAP-items) D_{pc}
4. The number and size of packets belonging to an WAP-item are conducted about the information about the TCP-protocols mechanisms and their influences and the distributions of
    - WAP-item sizes on Uplink S_{iu}
    - WAP-item sizes on Downlink S_{id}
5. The timing during a WAP-item is divided in five (WDP) or seven (WTP) parts to correspond to the model presented in figure 2
    - the transmission time of the Uplink packet (WAP-request, begin an item) D_{wu}
    - the processing time of WAP-request D_{pu}
    - the WWW-transaction waiting time D_{www}
    - the processing time of WAP-response D_{pd}
    - the transmission time of the Downlink packet (WAP-response) D_{wd}
    - the acknowledgement times on Uplink D_{au} and Downlink D_{ad}

- Presently there are distributions for the WAP-item size and WAP-transactions internal timings
- With $\boldsymbol{D_{wap}}$, $\boldsymbol{N_{pc}}$ and $\boldsymbol{D_{pc}}$ the problem is that IP address often changes during WAP-sessions, when GSM-data connection disconnects for idle periods. After that there in no information about the original user.
- a “browser-session” does not model users on the higher levels.
- 122 550 WAP/WWW-items are distributed to 11697 “browser-sessions” of which ~ 10 % do overlap and only ~60 % are separated by over 5 minute period.

Figure 4.1

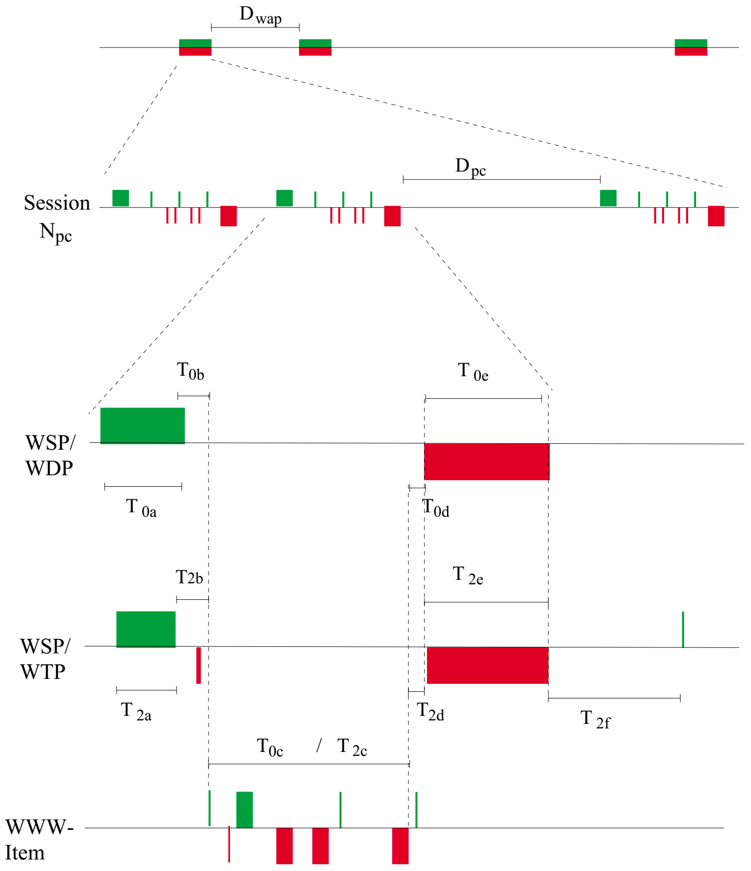

Figure 2. The model of WAP transaction timing.

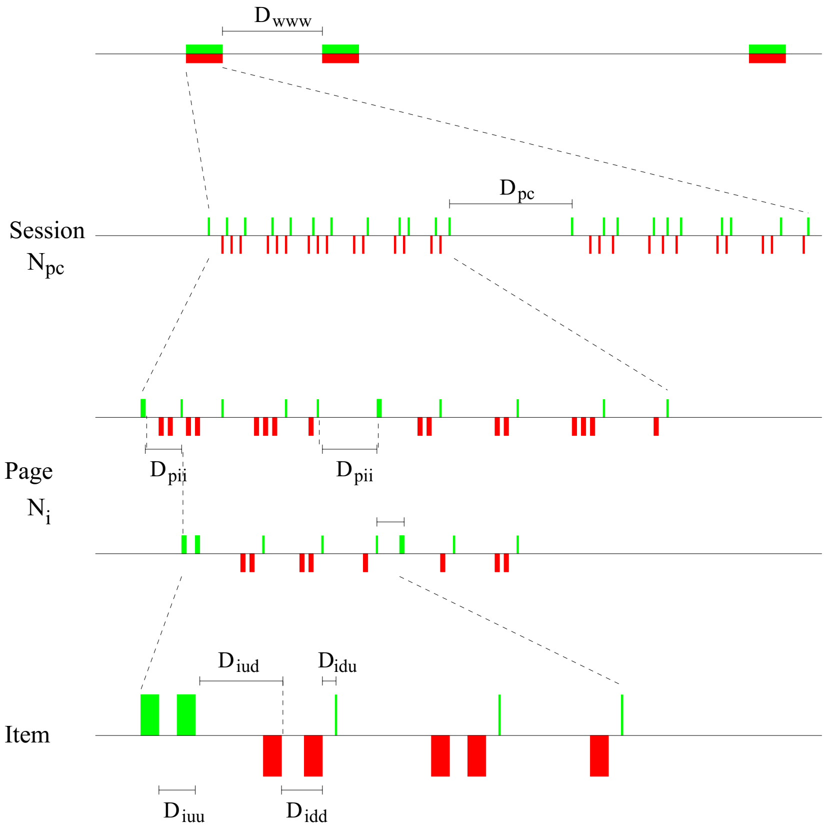

Figure 3. The model of WWW-session timing.

The ETSI Non-Real Time Traffic Model

- Referred later as ETSI model. Presented in D-ETR SMG-50402 v0.9.3: 4/1997, Annex 2, and enhanced in Technical Report TR 101 112 V3.2.0 (1998-04) Annex B, in both Paragraph 1.2.2. Traffic models / Non-real time services

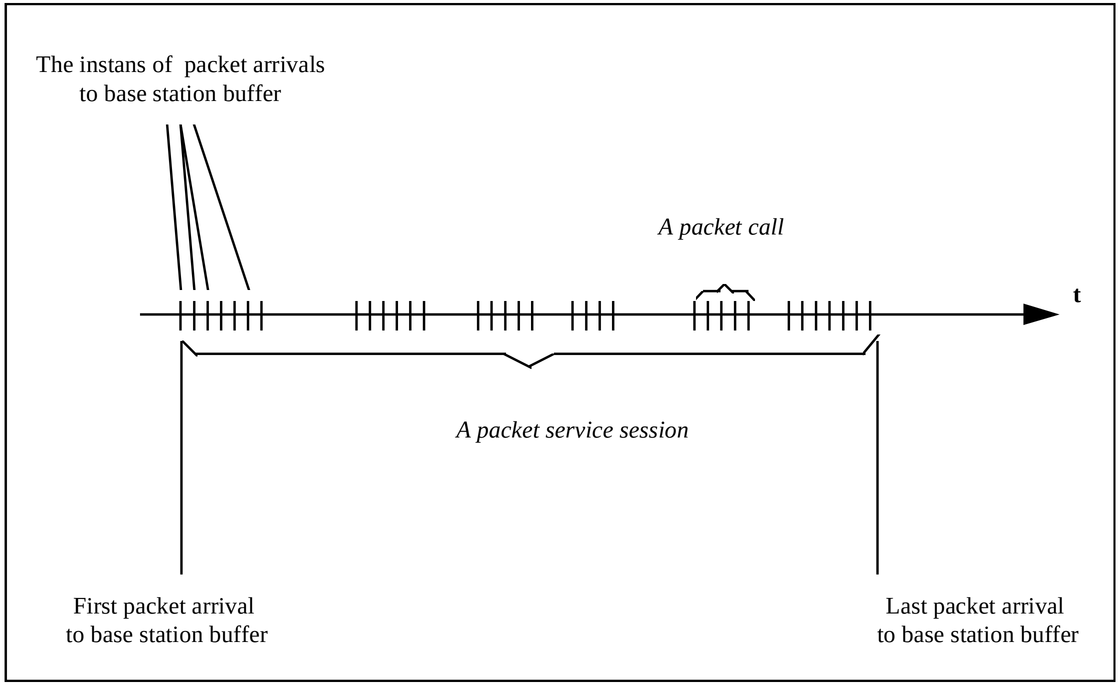

Figure 1.0. Typical characteristic of a packet service session.

- Figure 1.0 depicts a typical WWW browsing **session**, which consists of a sequence of **packet calls**.
- We only consider the packets from a source, which may be at either end of the link but not simultaneously.
- The user initiates a packet call when requesting an information entity.
- During a packet call several **packets** may be generated, which means that the packet call constitutes of a bursty sequence of packets.
- It is very important to take this phenomenon into account in the traffic model.
- The burstyness during the packet call is a characteristic feature of packet transmission in the fixed network.
- A packet service session contains one or several packet calls depending on the application. For example in a WWW browsing session a packet call corresponds the downloading of a WWW document.
- After the document is entirely arrived to the terminal, the user is consuming certain amount of time for studying the information. This time interval is called **reading time**.
- It is also possible that the session contains only one packet call. In fact this is the case for a file transfer (FTP). Hence, the following must be modeled in order to catch the typical behavior described in Figure 1.0:

- Session arrival process Modeled as a Poisson process. Has *nothing to do with call termination*.
- Number of packet calls per session, ***Npc*** $N_{pc} \in Geom(\mu_{Npc})$.
- Reading time between packet calls, ***Dpc*** $D_{pc} \in Geom(\mu_{Dpc})$
- Number of datagrams within a packet call***, Nd*** $N_d \in Geom(\mu_{Nd})$.
- Inter arrival time between datagrams (within a packet call) ***Dd*** $D_d \in Geom(\mu_{Dd})$.
- Size of a datagram***, Sd*** Pareto distribution is used

The session length is modeled implicitly by the number of events during the session.

Table 1.1 Characteristics of connection-less information types (default mean values for the distributions of typical www service)

Packet based information types	Average number of packet calls within a session	Average reading time between packet calls [s]	Average amount of packets within a packet call []	Average interarrival time between packets [s][1]	Parameters for packet size distribution
WWW surfing					
UDD 8 kbit/s	5	412	25	0.5	k = 81.5
UDD 32 kbit/s	5	412	25	0.125	α = 1.1
UDD 64 kbit/s	5	412	25	0.0625	
UDD 144 kbit/s	5	412	25	0.0277	
UDD 384 kbit/s	5	412	25	0.0104	
UDD 2048 kbit/s	5	412	25	0.00195	
(originally UDD 8 kbit/s)	5	12	15	0.96	

[1] The different interarrival times correspond to average bit rates of 8, 32, 64, 144, 384 and 2048 kbit/s.

According to the values for α and k in the Pareto distribution, the average packet size μ_n is 480 and average requested file-size is μ_{Nd} x μ = 25 x 480 bytes ≈ 12 kBytes. The packet size is limited to 66 666 bytes, giving a finite variance to the distribution. (First the truncations effect were neglected giving μ_n = 896 bytes and μ_{Nd} x μ = 15 x 896 bytes ≈ 13,4 kBytes.)

- The principle of dividing the model to layers like session, packet call and a packet is very good and describes the quite closely the actual process
- major drawback in the presented model are:

1. it does not take in to the consideration the direction of the packets
   - measured WWW traffic has great asymmetry
   - delays are different for example up to Down (~RTT) and down to up
   - used protocols can differ between Uplink and Downlink
2. WWW-pages are often composed of several (on average 4.8) WWW-items which use more than one parallel TCP-connections.
3. the systematic usage of selected statistic distributions can mask out some typical features.
   - For example the datagram (=packet) size and average interarrival time distributions.

The timing diagram presented in the figure 2. WAP transaction is there divided in following parts:

1. WAP-request transmitting time T_{0A}/T_{2A}. Calculated by dividing the packet size by line speed 9,6 kbit/s

2. WAP-request processing time in Gateway T_{0B}/T_{2B}

3. WWW-transaction waiting time T_{0C}/T_{2C}

4. WAP-response processing time T_{0D}/T_{2D}

5. WAP-response transmitting time T_{0E}/T_{2E}. Calculated by dividing the packet size by line speed 9,6 kbit/s.

6. WAP-response acknowledgement time T_{2F} (only in WTP). The time used by to the Mobile terminal to (process and) accept the WAP-response. The minimum = 26 ms. The measured from 32 ms to 12,6 s (mean 778 ms).

	WSP/WDP (WAP0)	WAP1 (= WAP2+rep)	WSP/WTP (WAP2)	WWW (WAP3)
Packets up	35 726	245 350	238 948	1 001 830
Packets down	35 831	297 241	242 609	940 535
Data-Packets up	35 726	123 288	122 550	137 550
Data-Packets down	35 831	175 996	121 366	321 312
IP-bytes up [kB]	4 838	14 777	14 508	95 673
IP-bytes down [kB]	17 287	66 681	56 262	153 376
Data-bytes up [kB]	3 802	7 044	6 974	55 035
Data-bytes down [kB]	16 248	57 466	48 740	179 741
Mean Item size up	136	136	136	136
Mean Item size down	479	479	479	479
Bursts	27 901	137 064	97 435	392 309
WAP/WWW-items	35 604	122 550	122 550	136 999
WSP/TCP-connections	35 604	122 651	122 651	138 299
WAP/WWW -pages	28 882	85 243	89 492	122 500
WAP/WWW-sessions	3 028	11 722	11 723	7 467
Burst time [s]	15 178	171 667	160 526	127 716
Item time [s]	52 856	404 943	128 039	78 789
TCP-connection time [s]	52 856	541 389	270 440	75 731 900
Page time [s]	46 614	491 371	269 959	77 703
Session time [s]	491 546	1 901 940	1 708 820	2 369 690

Table 1. The main statistics of data measured Packets, IP-bytes and Data-bytes, the mean sizes of WAP&WWW-Items, the numbers and total lengths of Bursts, Nibbles, WSP-connections, WAP&WWW-items, -pages and -sessions.

Measures for average Times	Means		Medians		Mean for models	
WSP/WDP	ms	%	ms	%	ms	%
WAP-request transmitting	113,36	6,9	112,00	20,0	113,30	10,9
WAP-request processing	24,80	1,5	2,00	0,4	6,07	0,6
WWW-transaction	541,59	32,8	79,40	14,2	453,72	43,6
WAP-response processing	591,33	35,8	50,10	9,0	73,62	7,1
WAP-response transmitting	381,38	23,1	316,00	56,5	393,74	37,8
Total	1652,46	100	559,50	100	1040,44	100
WSP/WTP	ms	%	ms	%	ms	%
WAP-request transmitting	75,23	6,7	63,10	10,1	75,81	5,6
WAP-request processing	22,37	2,0	2,51	0,4	19,24	1,4
WWW-transaction	469,48	41,8	141,00	22,6	451,04	33,4
WAP-response processing	101,94	9,1	20,00	3,2	352,89	26,1
WAP-response transmitting	454,17	40,4	398,00	63,7	452,98	33,5
[Acknowledgement from mobile]	777,72	69,2	708,00	113,4	764,21	56,5
Total	1123,19	100	624,61	100	1351,95	100

Table 2. The average times for different parts of WAP-transaction.

Measures for average Times	Means	Medians	Mean for models
WSP/WDP	s	s	s
WAP-Transaction duration	1,979	0,636	
WAP-Page duration	2,109	0,695	
WAP-session duration	162,828	71,295	
WAP-Transaction separation	881	5,815	
WAP-Page separation	1087	9,505	
WAP-session separation	10223	1119,505	
WSP/WTP	s	s	s
WAP-Transaction duration	2,280	1,335	
WAP-Page duration	3,092	1,855	
WAP-session duration	145,842	70,875	
WAP-Transaction separation	254	7,865	
WAP-Page separation	348	14,025	
WAP-session separation	2537	446,925	

Table 3. The average times for duration and separation for WAP-transactions, WAP-pages and WAP-sessions.

The activity during WAP-transactions and -sessions

The activity we defined as the minimum time needed to transfer the measured IP-packets over the given bandwidth

Activity during Transactions	By mean		By median		By model	
WDP-Transaction up	113,36	6,9	112,00	20,0	113,30	10,9
WDP-Transaction down	381,38	23,1	316,00	56,5	393,74	37,8
WDP-Transaction duration	1652,46		559,50		1040,44	
WTP-Transaction up	75,23	4,0	63,10	4,7	75,81	3,6
WTP-Transaction down	454,17	23,9	398,00	29,9	452,98	21,4
WTP-Transaction duration	1900,91		1332,61		2116,17	

Table 4. The Activity during WAP-transactions.

Activity during sessions	By mean		By median		By model	
WDP-Transaction duration	1,98	1,2	0,64	0,9		
WDP-Transaction number	10,76		10,76		10,76	
WDP-Transactions total	21,29	13,1	6,84	9,6		
WDP-session duration	162,83		71,29			
WTP-Transaction duration	2,28	1,6	1,34	1,9		
WTP-Transaction number	9,46		9,46		9,46	
WTP-Transactions total	21,58	14,8	12,63	17,8		
WTP-session duration	145,84		70,88			

Table 5. The Activity of WAP-transactions during WAP-sessions.

- The user activity during WAP-sessions will be a result from multiplication of
  - the activity factor inside WAP-transactions (Table 4) and
  - the part WAP-transactions take during the WAP-sessions (table 5).
  - WSP/WDP uses uplink 7- 20 % and WTP/WSP only 4-5 %.
  - WSP/WDP uses downlink 23- 57 % and WTP/WSP only 21-30 %.
  - The ratios between uplink and downlink are 1: 2,8-3,5 for WDP and about 1: 6 for WTP.
  - The relations between IP bytes transferred are 1: 3,6 for WDP and 1: 3,9 for WSP (incl. opening and closing).

- In matched transactions the WAP has compressed the data on average to 20 - 43 % compared to WWW
- The total relation of transferred bytes is 37 % with 92,9 Mb of WAP (WDP and WTP together) and 249 MB of WWW traffic.
- If an end-to-end WWW would be used the wireless link activity would increase 168 % and the times in table 2 would increase 0,9 - 1,2 seconds

- The WWW-items created by WAP are smaller and time intervals between WWW-packets are mostly larger than with normal WWW-items. Most request and responses fit to a single packet.
- Keep-alive packets should be excluded from all the statistics of WWW-items, pages and sessions.