PERIODICA POLYTECHNICA SER. EL. ENG. VOL. 44, NO. 1, PP. 51–64 (2000)

WIRELESS APPLICATION PROTOCOL PERFORMANCE

Lóránt FARKAS and Lajos NAGY

Department of Microwave Telecommunications
Budapest University of Technology and Economics
H–1521 Budapest, Hungary
e-mail: farkas@nov.mht.bme.hu, t-nagy@nov.mht.bme.hu

Received: 30 Nov. 2000

Abstract

Our work addresses the various performance testing aspects of the WAP architecture. For a systematic approach at first the architectural elements have to be identified. Then appropriate testing methods have to be found for testing each element and for the test of the entire architecture.

In this paper we shall outline a theoretical basis for testing the influence on the performance of the wireless network and the IP backbone through the service quality parameters. First the service quality parameters are introduced from the provider and the user viewpoint, then the effect of the network (wireless and IP backbone) through bearers and transport protocol overheads on these parameters is analyzed. Conclusions are drawn regarding the results of the analysis and further steps are outlined for the refinement of our study. The results have not been confirmed by measurements yet, we are working on appropriate test scenarios to validate the outcomes and to refine our model.

The final purpose of the work is to specify a performance parameter monitoring system for in-use WAP architectures that would tune the WAP parameters 'on the fly' in order to optimise the service quality parameters from the user viewpoint.

Keywords: service quality, user viewpoint, provider viewpoint, WAP, WTP, WSP, HTTP, wireless gateway, wireless network, IP backbone.

1. The WAP Architecture

The simplified architecture of a WAP system can be viewed as in *Fig. 1*.

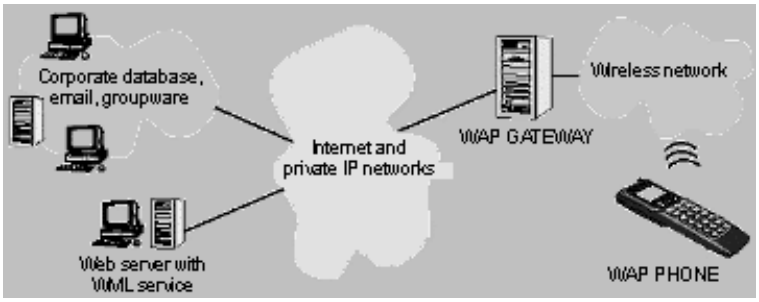

Fig. 1. WAP architecture

The WAP infrastructure is basically made of five elements: the client mobile phone, that implements a complete WAP stack; the wireless network, the communication medium between the client mobile phone and the peer; the wireless gateway, the peer of the client mobile phone, it implements another complete WAP stack, but it also behaves like a client from the IP stack point of view; the IP backbone, the medium between the wireless gateway seen as a client and the peer from the point of view of the IP protocol stack; the content provider, offering usually Wireless Markup Language (WML) contents to the wireless gateway seen as a client.

The different elements of the architecture influence separately the service quality parameters. The contribution of each element on each parameter has to be found.

2. Service Quality Parameters – the Different Viewpoints

The service quality parameters can be divided into measurable ones, that can also be very rigorously defined, and qualitative ones, important for the user, that are not necessarily very rigorously defined, nor directly measurable. The first kind of parameters can be called the Provider's or the network's, the second kind the User's or the client's. *Fig. 2* presents these different views.

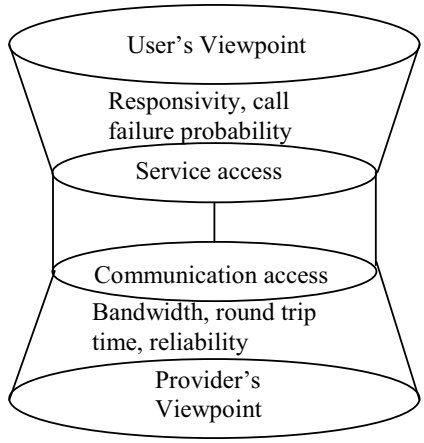

Fig. 2. Different viewpoints on the service quality

The third kind of parameters are the effective parameters of the WAP and TCP/IP stacks of the end systems and of the wireless gateway, these can be effectively tuned.

These two viewpoints can be considered appropriate in the case of WAP architectures, because we have measurable parameters on one hand and the user on the other hand, who does not seem to be affected by these parameters, but rather by

the responsiveness of his WAP phone. The main task would be to find appropriate measurement methods and strategies for the first kind of parameters, relationships between the two kinds and then based on these relationships, to find adaptive tuning strategies for the effective parameters in order for the service quality parameters from the user viewpoint to be optimised.

3. Service Quality Parameters – the Provider Viewpoint

End-to-end transit delay or *response time* is the elapsed time for a WDP/UDP datagram to be passed from the sender, through the network, to the receiver. From the provider viewpoint it is useful to consider this parameter separately for the lowest common denominator, which is the WDP/UDP datagram. No direct measurement methods are available as precise clock synchronization between the peers cannot be achieved.
Turnaround (round-trip) time (rtt) – easier to measure than the response time, because we do not need precise synchronization between the peer clocks. It is practically the double of the response time.
Jitter (j) – is the variation in end-to-end transit delay or response time or turnaround time.
Bandwidth (B) – is the maximal data transfer rate that can be sustained between two end points, measured from this viewpoint as transferred information bits/s.
Reliability (R) – average datagram loss probability.

4. Service Quality Parameters – the Client Viewpoint

Call setup time – the time perceived by a user between launching a data call and the first result that appears on the display of his mobile phone.
Call release time – the time perceived by a user between explicitly releasing the call and the result of his action appeared on the screen.
Call establishment success probability – the ratio of successfully established WAP data calls to the total number of WAP data call attempts launched by the user.
Call release success probability – the ratio of successfully terminated calls to the total number of initiated call releases.
Average transaction time (T) – the average time needed for a transaction to successfully complete.
Transaction failure probability (p) – the ratio of the failed transactions to the total number of requested transactions.

The analysis of the first four parameters is beyond the reach of this study, because they involve circuit switched bearers and they do not strictly belong to the WAP architecture, but rather to the underlying PPP and wireless physical link. The optimisation of the last two parameters is a real challenge and should be the

purpose of an adaptive parameter-tuning algorithm. In this paper we focus solely on the average transaction time.

5. A Simple Model of Mapping the Two Viewpoints

The response time can be expressed as follows:

$$T = rtt + \text{req}_{\text{WSP}} + \text{req}_{\text{HTTP}} + \text{processing} + \text{repl}_{\text{WSP}} + \text{repl}_{\text{HTTP}}. \quad (1)$$

In other words, it depends on the round-trip time (the sum of wireless and IP backbone impacts), on the processing through the network path and on the request and reply times, of the two end equipments (mobile client, content provider) and of the gateway. req_{WSP} is the time needed for the client to emit the request onto the wireless network and rep_{WSP} is the time needed for the gateway to emit the reply onto the wireless network. req_{HTTP} is the time needed for the gateway to emit the request on the IP backbone and finally $\text{repl}_{\text{HTTP}}$ the minimum amount of time needed by the content provider to emit the reply on the IP backbone:

$$\text{req}_{\text{WSP}} = \text{req}_{\text{WSP}}/B_{\text{wireless}}, \quad (2)$$

$$\text{repl}_{\text{WSP}} = \text{repl}_{\text{WSP}}/B_{\text{wireless}}, \quad (3)$$

$$\text{req}_{\text{HTTP}} = \text{req}_{\text{HTTP}}/B_{\text{IP}}, \quad (4)$$

$$\text{repl}_{\text{HTTP}} = \text{repl}_{\text{HTTP}}/B_{\text{IP}}. \quad (5)$$

rtt is the round-trip time, B_{IP} and B_{wireless} the bandwidths of the IP backbone and the wireless network.

From (1) it can be seen that except the term 'processing' the other terms are related to the physical link, therefore they cannot be optimized, once the physical link is given. The term 'processing' can be further divided into terms:

$$\text{processing} = \text{overhead}_{\text{wireless}} + \text{overhead}_{\text{IP}} + \text{delay}_{\text{GW}} + \text{delay}_{\text{server}}. \quad (6)$$

All four terms have to be carefully examined. The terms depend themselves on the service quality parameters from the network viewpoint and also on the parameters of the TCP/IP and the WAP stack. The delay introduced by the server depends on the current load on the server. On the gateway side it depends on the load (through the sizes of the queues) and on the efficiency of the protocol conversion algorithm between the two stacks.

The four terms can be further expanded as follows:

$$\text{overhead}_{\text{wireless}} = \text{overhead}_{\text{WTP}}, \quad (7)$$

$$\text{delay}_{\text{GW}} = f_1(\text{queue size, protocol and code conversion}), \quad (8)$$

$$\text{delay}_{\text{server}} = f_2(\text{load}), \quad (9)$$

$$\text{overhead}_{\text{IP}} = \text{overhead}_{\text{TCP}}. \quad (10)$$

Finally, we have the five parameters, on the right sides of (7), (8), (9) and (10), to be evaluated and minimized. We propose at this stage to analyze the effect of the protocol overheads, from *Eqs.* (7) and (10). The effect of the gateway and the content provider constitute the subject of a further study.

For the simplicity of the study we will assume, that a complete transaction is made of two separate transactions: a WSP transaction between the mobile client and the gateway and an HTTP transaction between the gateway and the content provider.

6. Some Characteristics of the WAP Traffic

At the moment it can be supposed to be of browsing, request-response type. Therefore the same methodology, of dividing the WML contents into classes, could be followed, as in HTML. So far, to our knowledge, there does not exist an estimation or proposal based on measurements for WML length distribution or classes on the Internet. From a number of 100 randomly chosen WML pages from different sites, as a result of POST and GET operations, we obtained the following outcome:

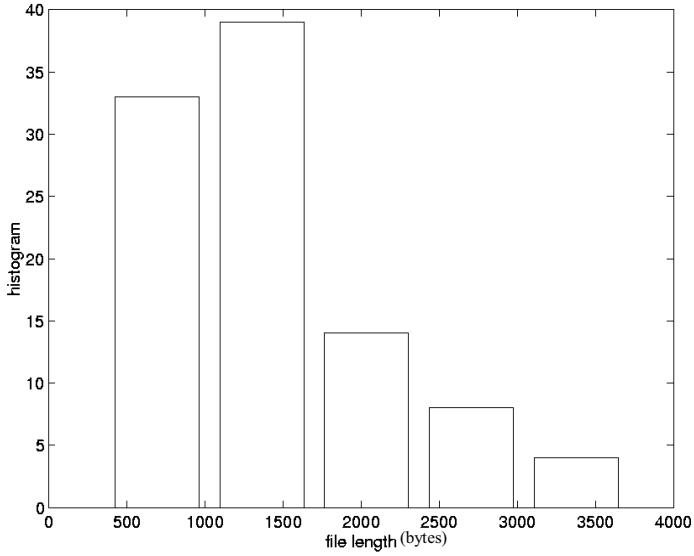

Fig. 3. WML page length distribution

A classification based on the logarithm of average file lengths is not meaningful, since the size of the files seldom reaches 5000 bytes of length, therefore a linear scaling can be proposed, as in *Fig. 3*, into 5 classes. It is probably not very accurate, further refinements will be necessary as our knowledge about WML traffic

expands and the number of available server logs grows. However, we consider that the figure is meaningful at least from a qualitative viewpoint.

7. The Effect of TCP

The TCP protocol adds its impact through two mechanisms: the three-way handshaking, during the connection opening procedure, and the slow start procedure, during the bandwidth evaluation phase. This impact can be quantified as follows:

$$\text{overhead}_{\text{TCP}} = rtt_{\text{IP}} + t_{\text{slowstart}} + \text{losses}. \tag{11}$$

These effects generally have a large impact on the transfer of small files, because TCP has not been designed for request-response transfers of small WML pages and the session probably would not even come out of the slow start procedure, when the transfer will already have been finished. Therefore the three-way handshaking and the slow start procedure have a considerable overhead on the average transmission time of small files.

Typical round-trip time and bandwidth values for different IP backbone bearers are given in *Table 1* [2]:

Table 1. Typical values for rtt and B

Network	rtt	Bandwidth
Ethernet	0.7 ms	8.72 Mbit/s
Fast Ethernet	0.7 ms	100 Mbit/s
Slow Internet (between different continents)	161 ms	0.102 Mbit/s
Fast Internet (sites on the same continent)	89 ms	1.02 Mbit/s
Modem	250 ms	0.0275 Mbit/s
ISDN	30 ms	0.122 Mbit/s
ADSL	30 ms	6 Mbit/s

The slow start effect depends on the used TCP/IP implementation.

It is useful to consider separately the idealistic transaction time on the IP backbone and on the wireless network separately. Therefore the overhead of TCP on the IP backbone-located part of the transaction is examined.

$$T_{\text{IP min}} = rtt_{\text{IP}} + \text{req}_{\text{HTTP}} + \text{processing}_{\text{HTTP}} + \text{repl}_{\text{HTTP}}. \tag{12}$$

Ideally, the 'processing' term is 0. In a realistic case, when processing is neglected but the effects of TCP are considered, the realistic value will be:

$$T_{\text{TCP}} = 2 \cdot rtt + \text{req}_{\text{HTTP}} + t_{\text{slowstart}} + t_{\text{delayed\_ack}} + \text{repl}_{\text{min}}. \tag{13}$$

TCP request pipelining could be considered at the gateway side if the client were an intelligent browser that requested in parallel more than one content from the provider. In this case the extra rtt would not have mattered if n requests had been pipelined – the overhead would have been thus rtt/n. However, for WAP this is not the case, the microbrowser will presumably request one file at time.

The maximum size of the data in a TCP segment is 536 bytes, therefore the transfer of a WML page will take usually more than 1 segment. So the slow start and/or the delayed acknowledge overhead should be taken into consideration.

[2] identifies three different ways of congestion window openings and acknowledgement policies in modern TCP implementations, as shown in *Table 2*:

Table 2. Number of segments (simple and accumulated) between stalls during slow start procedure for 3 different policies (seg – segment number, ac_seg accumulated segment number)

Stalls	No delayed ACK policy	Delayed ACK policy	ACK every segment policy
1	2 (2)	2 (2)	2 (2)
2	3 (5)	3 (5)	4 (6)
3	3 (8)	5 (10)	8 (14)
4	6 (14)	8 (18)	16 (30)
5	9 (23)	12 (30)	32 (62)
6	12 (35)	18 (48)	64 (126)
7	18 (53)	27 (75)	128 (254)
8	27 (80)	41 (116)	256 (510)
9	42 (122)	62 (178)	512 (1022)
10	63 (185)	93 (271)	1024 (2046)

For the slow start overhead evaluation the following formula can be used [2]:

$$\text{slowstart} = \sum_{i=1}^{\text{nbr.stalls}} \left(rtt - \frac{(\text{seg}(i) - k) \cdot \text{segm.size}}{B_{\text{IP}}} \right), \tag{14}$$

where $k = 1$ or 2 if every segment is acknowledged (3rd column) or delayed acknowledge is used (1st and 2nd columns).

In the evaluation we will neglect the time needed to issue the request. Therefore only two terms remain: the round-trip time and the time needed to put the server response, that is the WML content on the link.

Figs. 4, *5* and *6* present the overhead of TCP as a function of rtt and bandwidth, for a WML file length of 2500 bytes (TCP data segment size has been considered to be 512 bytes):

It can be observed that in the worst case for an average file length the overhead is less than 3 times the idealistic transaction time for the IP backbone.

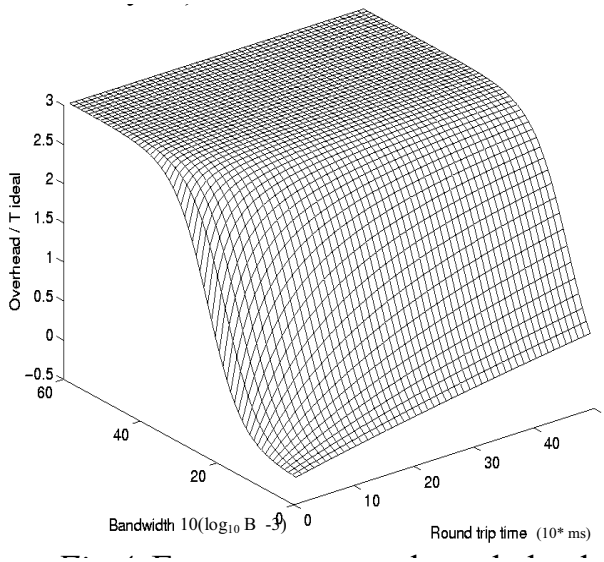

Fig. 4. Every segment acknowledged (3rd column)

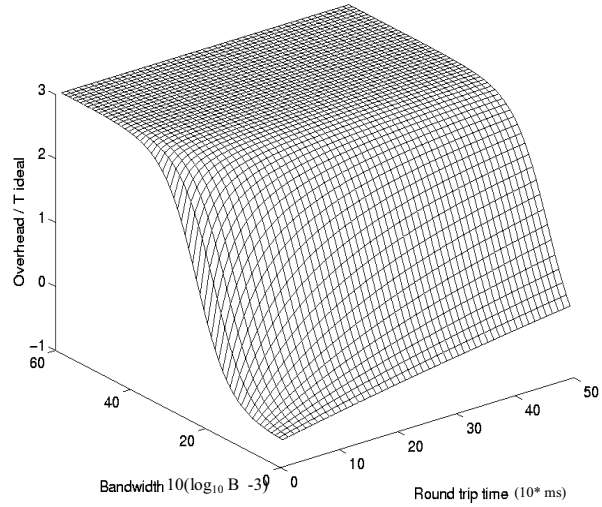

Fig. 5. Delayed acknowledge (columns 1 and 2)

8. The Effect of WTP and WSP

The WTP lays on the top of WDP, a non-reliable transport protocol equivalent to UDP. It contains the additional features that make of the transport layer of WAP a

reliable kind. It has been optimized to service WSP, the equivalent of HTTP. It is basically a request-response transport protocol. In the following we will study the effect of the connection-oriented service of WSP that relies on the services of the WTP layer.

The basic WSP transaction, called method invocation, is shown in *Fig. 6* [5]:

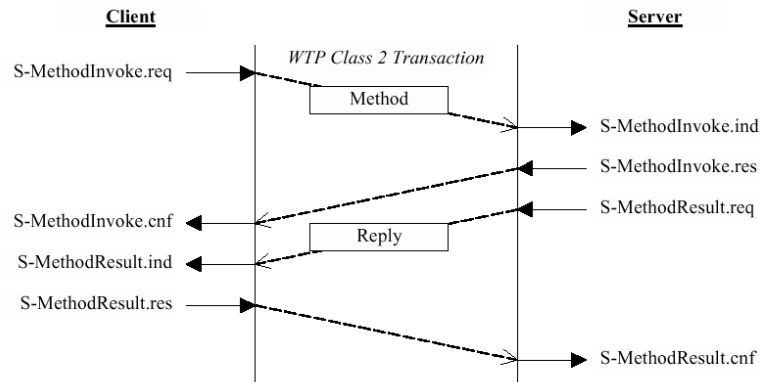

Fig. 6. WSP method invocation

It is presumed that ordinary method invocation would be the most common WSP operation throughout the lifetime of a WSP session.

The basic class 2 transaction of WTP is shown in *Fig. 7* [6]:

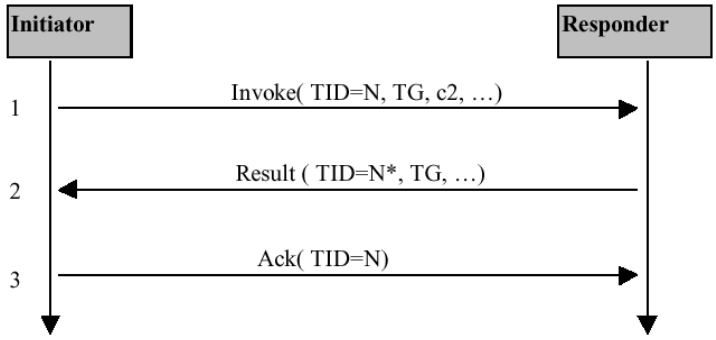

Fig. 7. Basic class 2 WTP transaction

(13) quantifies the overhead of the WTP on the transaction time:

$$\text{overhead}_{\text{wireless}} \approx 5 \cdot rtt_{\text{wireless}} + 2 \cdot \text{inv}_{\text{WTP}} + 4 \cdot \text{res}_{\text{WTP}} + 4 \cdot \text{ack}_{\text{WTP}} + \text{losses}. \quad (15)$$

The WAP Forum recommends values for the round-trip time for different bearer types, these are the following [6].

Table 3. Recommended values for median round-trip times and bandwidth

Bearer	Median round-trip time	Bandwidth
SMS	10 s	0.5 kbit/s
USSD	3 s	1.5 kbit/s
IP	0.2 s	7.2 kbit/s

These values, however, should be further studied. The bandwidth for the different bearers are highly dependent on the current load, the presented values are rather optimistic. If accepted, it means for example that there is no possible way of downloading a WML page using the SMS bearer in a shorter time than 1 minute (6 round-trip times).

The WTP claims to support wireless applications, because it does not include connection establishment and tear down processes. But it includes in every transaction three kinds of service primitives, the equivalent of 1.5 round-trip times, that in the case of SMS severely restricts the service quality.

The other terms in *Eq.* (14) can be expressed as follows:

$$\mathrm{inv}_{\mathrm{WTP}} = \mathrm{inv}_{\mathrm{size}}/B_{\mathrm{wireless}}, \tag{16}$$

$$\mathrm{res}_{\mathrm{WTP}} = \mathrm{res}_{\mathrm{size}}/B_{\mathrm{wireless}}, \tag{17}$$

$$\mathrm{ack}_{\mathrm{WTP}} = \mathrm{ack}_{\mathrm{size}}/B_{\mathrm{wireless}}, \tag{18}$$

where inv, res and ack represents the time needed to issue an invoke, a response and an acknowledge PDU on the wireless bearer, B_{wireless} is the estimated bandwidth of the bearer. The size of these PDUs is the following:

inv: 4 bytes,
res: 3 bytes,
ack: 3 bytes.

These PDUs are in addition to the data-carrying PDUs, therefore only the header size should be considered, the information part is missing.

It is useful to consider separately the idealistic transaction time on the IP backbone and on the wireless network. Therefore the overhead of WTP on the wireless network-located part of the total transaction time (19) is examined.

$$T_{\mathrm{wireless}} = rtt_{\mathrm{WTP}} + \mathrm{req}_{\mathrm{WTP}} + \mathrm{processing}_{\mathrm{WTP}} + \mathrm{repl}_{\mathrm{WTP}}. \tag{19}$$

Figs. 8–10 present the overhead/ideal response time ratio for the round-trip times characteristic for the three considered bearers. The size of the request PDU has been chosen of 32 bytes, the response PDU has been a parameter, equal to the length of the requested WML content, the wireless bandwidth has also been a parameter. We considered that the request and response fits into one PDU in each transfer or

alternatively that the WTP layer provides the segmentation and reassembly function. Otherwise the overhead increases with 1.5 rtt for each new segment to be transferred, because each will start a new class 2 WTP transaction.

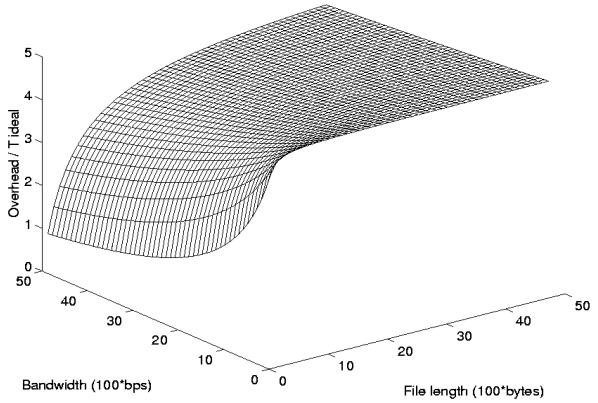

Fig. 8. WTP overhead, $rtt = 10$ s

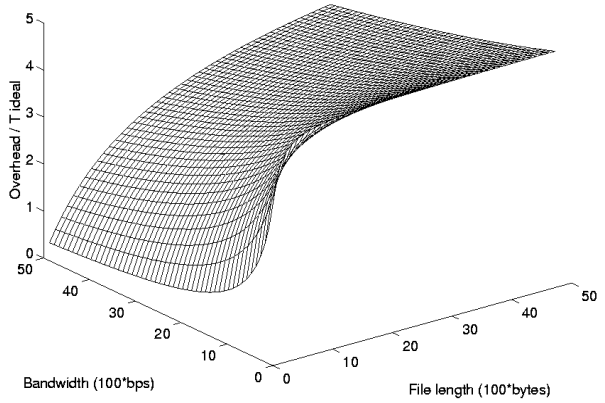

Fig. 9. WTP overhead, $rtt = 3$ s

It can be observed, that at high values of the wireless round-trip time the overhead grows quickly with the content length and then reaches a steady value, and it becomes also independent of the bandwidth. At medium round-trip time, the overhead strongly depends on both of the bandwidth and the file length. On the other hand, at small values for the round-trip time the overhead becomes insensitive on the bandwidth except very low bandwidths and file length, it seems that WTP optimizes this situation and does not behave well for long round-trip times.

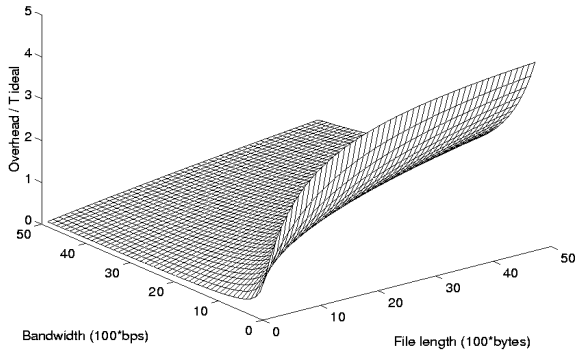

Fig. 10. WTP overhead, $rtt = 0.2$ s

Representing the overhead as a function of bandwidth and round-trip time gives us another insight:

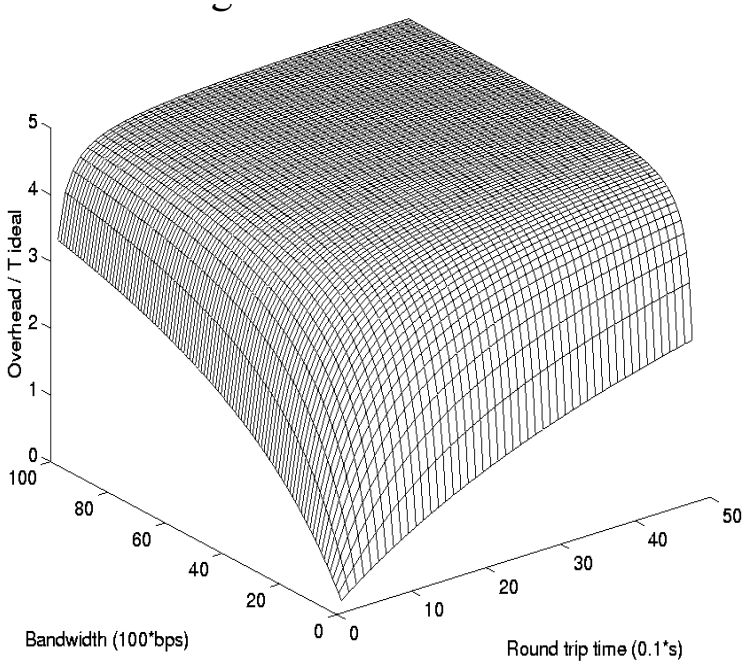

Fig. 11. WTP overhead, 500 bytes file length

From *Fig. 11* and *12* it results that with increasing length of the WML content the overhead decreases, except the natural growing dependence on the round-trip time and on the bandwidth.

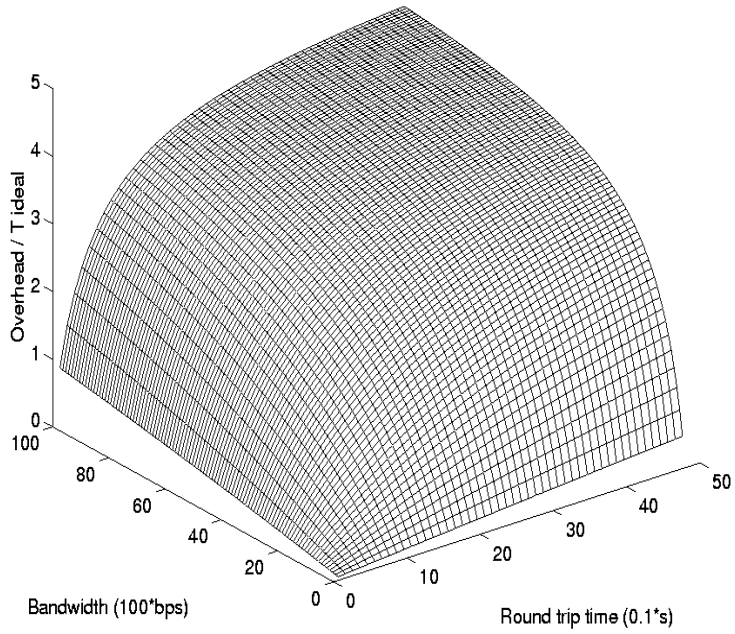

Fig. 12. WTP overhead, 5000 bytes file length

9. Conclusions and Further Steps

The impact on the overall transaction time is much greater on the wireless side of the communication. A 5-times overhead in this case is much more meaningful than a 3-times overhead on the IP backbone side, at least for the case of SMS and USSD bearers. Therefore in the evaluation of the overall transaction time in this case the effect of TCP may not be included. However, it should be noted, that we assumed on both sides of the communications error-free data transfer. Presumably the errors on the IP backbone will strongly affect the transaction time on the wireless network, because time-out conditions will occur and several other WTP class 2 transactions will stem in order to deal with these time-outs.

For the case of CSD type bearers (through modem) it can be observed that the round-trip times and therefore the transaction times become comparable if on the IP backbone side of the communication there is a weak link: modem or slow Internet. In this case the errors on each side of the network become equally important.

The further steps should focus on analyzing the impact of the errors handled at the TCP and WTP level on the overall transaction times. From this analysis we should be able to find strategies of tuning the WAP parameters inside the wireless gateway in order to minimize the number of additional WTP transactions, with a direct impact on the transaction time.

References

[1] HALSALL, F., *Data Communications, Computer Networks and Open Systems*, Addison-Wesley, 1992, pp. 512–521.
[2] HEIDEMANN, J. – OBRACZKA, K. – TOUCH, J., Modeling the Performance of HTTP Over Several Transport Protocols, *IEEE/ACM Transactions on Networking*, **5** No. 5, October 1997, pp. 616–630.
[3] STALLINGS, W., *Data and Computer Communications*, Prentice Hall, 1994, pp. 578–586.
[4] Wireless Application Protocol, Wireless Datagram Protocol Specification, 1999, www.wapforum.org.
[5] Wireless Application Protocol, Wireless Session Protocol Specification, 1999, www.wapforum.org.
[6] Wireless Application Protocol, Wireless Transaction Protocol Specification, 1999, www.wapforum.org.