

QoS estimation for cellular packet data networks

Nicolae Cotanis

R&D Department, LCC Intl., Inc.
McLean, VA, 22180, USA
e-mail: nicolae_cotanis@lcc.com

Abstract— There is a lot of effort in the industry for defining adequate and realistic end-to-end QoS indicators for cellular data services, which might be applicable for estimating delivered QoS and for specifying the QoS control mechanisms for the underline transport networks. The paper proposes a set of measurements and a procedure for mapping the results to key quality and performance indicators. The procedure is independent of the network architecture and technology. It is a top down approach for directing the network optimization process, starting from the end user perspective.

End-to-end QoS; cellular data services; CDMA 2000, 1xRTT, GPRS, resource allocation, traffic classes

I. INTRODUCTION

Despite increased consideration for end-to-end performance required by user applications, the term QoS is usually not well defined, or used loosely. QoS comes into attention when planning and deploying networks, as well as when monitoring service quality. Aspects of networks and services provisioning are presented in [1], but the standard is not application-oriented, and, in many areas, too vague for practical use. Nevertheless, [1] sets the QoS framework and states the widely used definition as “the collective effect of service performance which determines the degree of satisfaction of a user for this service”.

Service functions (e.g. service management, connection quality, billing, customer net/serv. management) and *service quality* criteria (e.g. speed, accuracy, availability, reliability, security, simplicity, flexibility) are considered for creating a *QoS matrix* for each service [2]. The paper elaborates on testing the connection quality function and its relevant QoS criteria such as *speed*, *accuracy*, *availability*, and *reliability*.

Considering the goals and the achievements for each party (customer and provider) involved in the service, one may devise four complementary viewpoints: customers’ QoS requirements / perception, and service provider QoS offering / achievement. Customer requirements are important when creating a *QoS test plan* for estimating the QoS delivered by a service provider. These requirements are usually expressed in non-technical language and focus on user-perceived effects, rather than their causes within the network. The delivered QoS is expressed in values assigned to *QoS indicators*, which are used for tracking performance and directing optimization.

The paper presents the concepts for estimating the delivered quality of cellular data services, and shows performance results of present day cellular packet data networks. Using basic application protocols, performance parameters are measured and mapped to delivered quality indicators.

Section II introduces the traffic classes as the starting point for QoS testing and anticipates the major characteristics of data services over cellular networks. Section III emphasizes on the test design and selection of applications. Section IV defines the quality and performance indicators, actual performance measurements for current CDMA 2000 and GPRS networks being given in Section V.

II. THE CELLULAR ENVIRONMENT

Evaluating the performance of a data network by checking each individual service offered to its subscribers might turn into a daunting task. One way to overcome such a challenge is to organize individual services into four basic traffic classes [4], [7]. Table 1 lists the traffic classes and their transport requirements. For selected data services, acceptable performance ranges for delay, delay variation and information loss are presented in [3], but they are too demanding for current cellular technology. The alternative would be to use as a baseline the average performance delivered by dial up landline connections, which stands as a de facto standard, representing the minimum performance subscribers would expect.

TABLE I. TRAFFIC CLASSES

Traffic Class	Applications	Delay	Jitter	Losses	Bit Rate
Conversational (Real time)	Voice, one to one video	S ^a	S		G ^d
Streaming (Real time)	Broadcast Audio Video	C ^b	C		G
Interactive (Best effort)	Web, Database, Games	L ^c		Low	
Background (Best effort)	E-mails, File transfers			Low	

a. Stringent, b. Constrained, c. Loose, d. Guaranteed

The end-to-end performance of a cellular data network is the result of the performance of all underlying bearers (Fig. 1). Changes in the propagation environment and/or network traffic load will result in major changes in the resources allocated to a

terminal during a call or from call to call. The Gaussian distribution of performance measurements, inherent for non-variable radio environments and traffic load, makes room to highly skewed distributions.

Starting from 2.5G, wireless networks allocate resources based on radio channel condition, traffic load, requested QoS, etc. Thus, for the same network traffic load and service, the delivered QoS (e.g. throughput or session time) will depend on the terminal- base transceiver station (BTS) distance. Terminals in the near range may get high data rate radio channels in comparison with terminals in the far range. Also, error control mechanisms, resource allocation and mobility management will further determine the delivered QoS.

The delivered QoS test must be designed for coping with such variability still offering meaningful and consistent results. Based on these, statistically significant problems are identified and faulty layers optimized in a top-down approach. While end-to-end QoS measurements are technology agnostic, measurements on radio access or core network layers demand specialized, technology dependent measurement tools (Fig. 1).

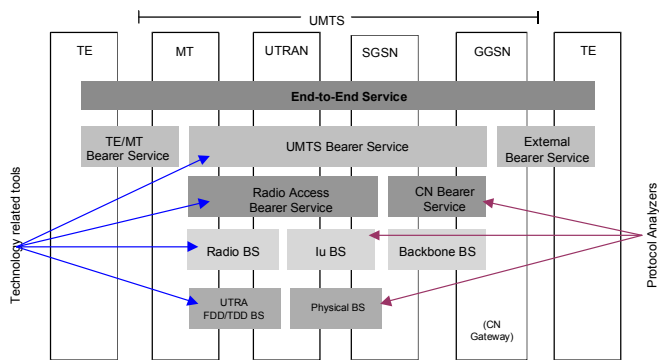


Figure 1. UMTS bearers [4]

III. DELIVERED QUALITY TEST DESIGN

When measuring delivered QoS for cellular networks, samples must be collected from locations uniformly distributed within the coverage area. Because the delivered QoS strongly depends on terminals' mobility, separate tests must be devised for stationary and mobile scenarios. The test design includes the selection of the Internet address for the server used in the process, the structure of the data session, and the driven routes. For measuring the performance of the cellular data network only, the test server must be directly connected to the network gateway; thus avoiding delays and congestions of the Internet.

A data session is defined as a set of applications executed by the test terminal. The data session begins by connecting the terminal to the data network (initial connection), continues by playing a specified set of services (e.g. PING, FTP get, FTP put, HTTP, etc), and ends by disconnecting from the data network (releasing the data network resources). The structure of the data session is designed based on the test goals.

The currently offered cellular data services belong to the background and interactive traffic classes (Table 1) being supported by "best-effort" IP techniques. Using a large set of

service types within a test data session dramatically reduces the spatial sampling resolution. We propose a method of estimating the basic quality indicators speed, accessibility, reliability and responsiveness based on two types of applications only: PING and FTP.

PING provides significant information regarding the stimulus response delay and IP network performance. PING messages are used for collecting statistics on Round Trip Time delay (RTT) as well as IP packet loss rate (PLR). Besides channel rate, RTT controls the offered throughput; high throughput requiring low RTT. The responsiveness quality is derived from RTT delays.

FTP or HTTP applications could be used for estimating uplink/downlink throughput performance, the old FTP having the advantage of avoiding measurement errors from file compression, cache, etc. Different payloads should be considered for quantifying the operation of the resource location algorithms. Throughput is used as a measure of speed quality. In addition, the success of the FTP sessions measures the reliability of the network as experienced by users.

FTP sessions require two communication ports. Usually, port 21 is used by the protocol interpreter (PI), while port 20 is used by the data transfer protocol (DTP). Port 21 sets up the data link (port 20) for performing the actual file transfer and closes the FTP session.

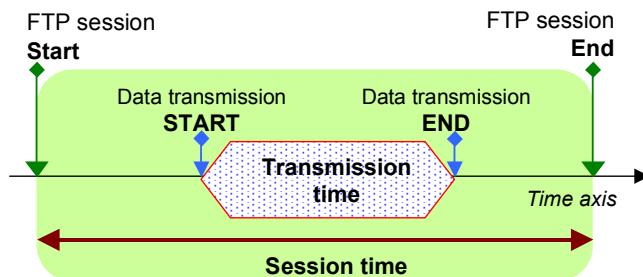


Figure 2. FTP session time structure

Based on the reference time used (1), one may derive FTP session and transmission throughput. The first one has significance for users while the second one gives indications on the network performance and may be used for directing the optimization process.

$$Throughput [kb/s] = \frac{Payload [kbit]}{Reference\ time [sec]} \quad (1)$$

IV. PERFORMANCE AND QUALITY INDICATORS

Table 2 gives the quality indicators, the performance parameters from which they are derived, and the data application exercised for estimating the performance indicators. The quality indicators copy the QoS criteria presented in [2]. *Accessibility* is related to service *availability* but it is customized for describing the initial access to cellular data services, when network equipment other than the one used to support the data services are used. *Responsiveness* was

included as a measure for quantifying the stimulus response delay that is important for interactive traffic classes.

TABLE II. QUALITY INDICATORS AND RELATED PERFORMANCE INDICATORS

Quality indicator	Performance parameter	Acronym	Unit	Application
Accessibility (A)	Connection success rate	CSR	-	Connect
	Connections set-up time	CT	second	
Speed (S)	Session throughput	ST	kb/s	FTP get/put
	Transmission throughput	TT	kb/s	
Reliability (RL)	FTP session success rate	SSR	-	
Responsiveness (RS)	Round trip delay	RTT	millisecond	PING

Quality indicators are dimensionless. They are derived from the cumulative distribution function (CDF) [6] of the performance indicators for specified values, representing *acceptable performance*. Equations (2)-(5) give the definitions for accessibility, responsiveness, speed, and reliability, respectively. Acceptable performance for indicator (*) is denoted by T(*).

Reliability is defined as the statistical frequency of sessions that completed successfully. Quality performance indicators may be combined using a weighted sum for providing a global figure for delivered QoS.

$$A = CDF_{CT}(T_A) \quad (2)$$

$$RS = CDF_{RTT}(T_{RTT}) \quad (3)$$

$$S = CDF_{Throughput}(T_{Throughput}) \quad (4)$$

$$RL = \frac{FTP_{UL/DL} \text{ successful}}{FTP_{UL/DL} \text{ requests}} \quad (5)$$

For cellular data services acceptable performances are not defined yet. The performance of data services over dialup landline POST connections could be considered acceptable performance for cellular packet data networks. For example, measurements made on FTP sessions over dialup landlines (Fig. 3) have indicated that 99.25% of the sessions required less than 11 seconds for transferring 10,000 byte files, which translates to a minimum average session throughput of 7.27 kbps. Therefore, 7.27 kbps could be considered as acceptable performance when calculating the delivered speed quality.

V. DELIVERED QOS RESULTS

Measurements were performed on GPRS and CDMA 2000 (1xRTT) networks for stationary and mobile terminals in order to collect performance statistics on connection time (CT), throughput and round trip time (RTT), which are used for deriving the quality indicators defined above. In addition,

mobile measurements (marked A-F) were made on multiple 1xRTT networks. Thus, performance variations due to technology and deployment were analyzed.

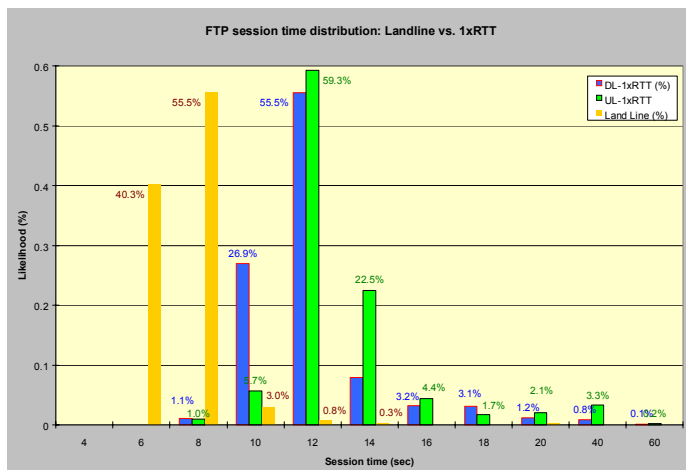


Figure 3. FTP session time statistics

A. Round trip time

Fig. 4 illustrates the RTT variation between technologies and with the terminal mobility. For a given technology, RTT acceptable performance may be derived from measurements on stationary terminals. For the same technology and mobility type (Fig. 5), a shift to the right in the RTT distribution or a large spread indicates poor network quality.

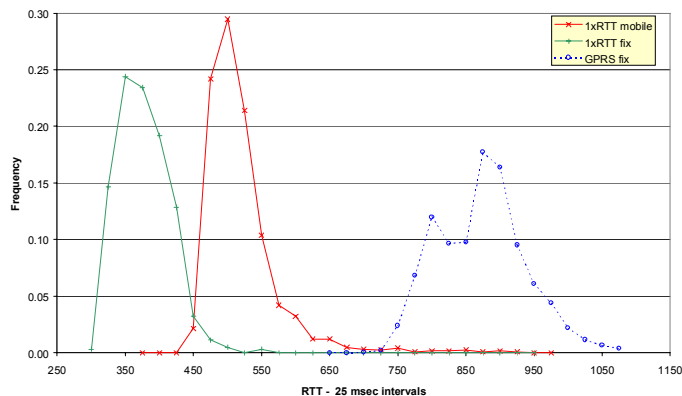


Figure 4. Round trip time distribution for GPRS and 1xRTT technologies

B. Application layer throughput

Transmission throughput is not directly perceived by users but provides valuable information on the performance of all the layers supporting a cellular data connection. The acceptable performance ($T_{Throughput}$) must be customized per technology, mobility type, and payload. Fig. 6 shows throughput cumulative distributions for stationary (1xRTT and GPRS) and mobile users in different 1xRTT networks.

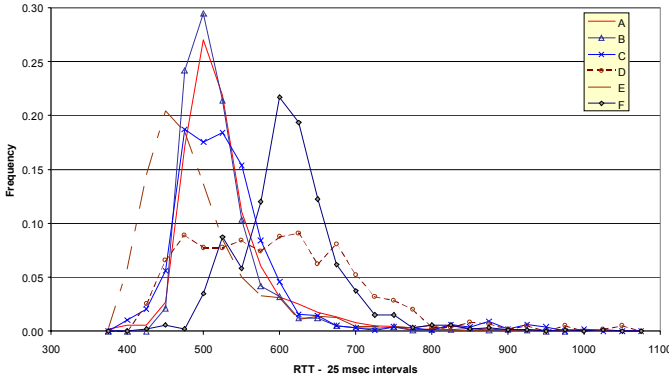


Figure 5. Round trip time distribution for different deployments (1xRTT)

A test plan including a non-homogeneous network configuration (macro- and micro-cells) results in a throughput distribution (Fig. 7) that generates non-consistent speed quality estimates (4). The low throughput figures identify sessions/locations where low data rate channels were allocated and many retransmissions happened, or, where channel resources were not available and the terminal moved frequently in the idle/control mode. The high throughput figures give an indication of the maximum performance that may be achieved in the network's micro-cells.

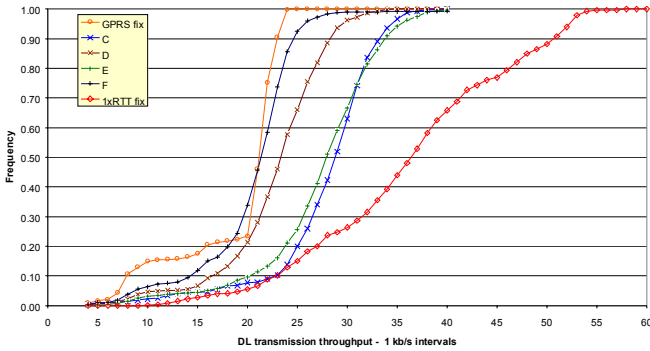


Figure 6. DL transmission throughput (10kB payload)

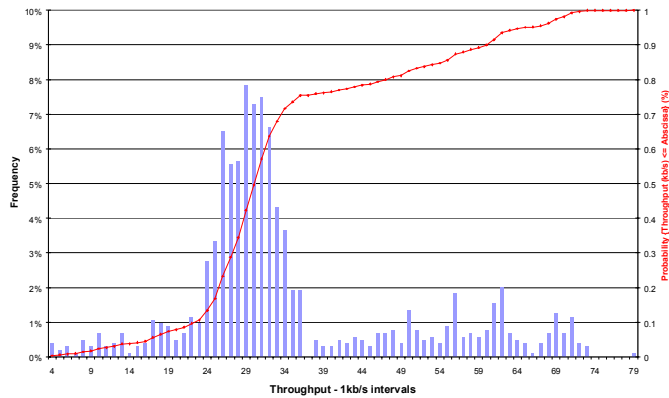


Figure 7. Transmission throughput distribution (10kB payload)

Even for a well-designed test plan, the throughput distribution is highly skewed; recommending percentiles and not averages to be used for deriving performance figures. For example, the available bandwidth performance is defined as the 90th-percentile of the throughput (6).

$$0.9 = \Pr\{\text{Throughput} \geq \text{Available bandwidth}\} \quad (6)$$

C. Performance of the resource allocation algorithms

The available bandwidth and/or the transmission time for different payloads may be used for assessing the overall performance of the resource allocation algorithms. In Fig 8, increasing the payload four times (from 10,000 bytes to 40,000 bytes), the 90th-percentile transmission time increases approximately 2.52 times (from 5.61 seconds to 14.16 seconds) as a result of higher data rate channels allocated for the larger payloads. For small payloads, a 10-fold increase in the file/page size, from 1024 to 10,000 bytes, corresponds to a 2.67-times increase in the 90th-percentile transmission time (from 2.10 seconds to 5.61 seconds). Optimization is required when the increase in the transmission time almost mirror the increase in the payload.

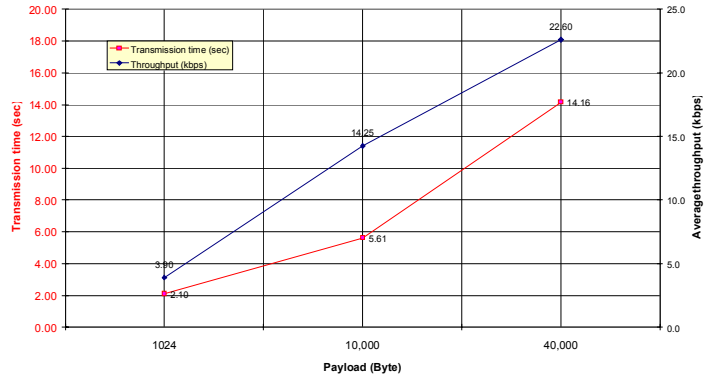


Figure 8. Available bandwidth as a function of the file size (1xRTT deployment)

D. IP network performance

The IP layer performance is expected to control the performance of the communication layers sitting on it. The bandwidth for TCP connections [5] depends on the average RTT and packet error rate p (7).

$$B(p) = \frac{const}{RTT \sqrt{p}} \quad (7)$$

For networks A-F, Fig. 9 and Fig. 10 show measured relationships for bandwidth-average RTT and bandwidth- p , respectively.

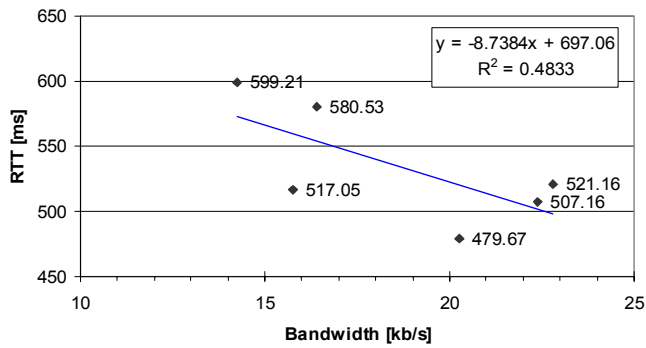


Figure 9. RTT versus bandwidth, networks A-F

The large errors in the least mean square estimation (R^2), do not recommend IP measurement as the only way for assessing the quality of cellular packet data networks.

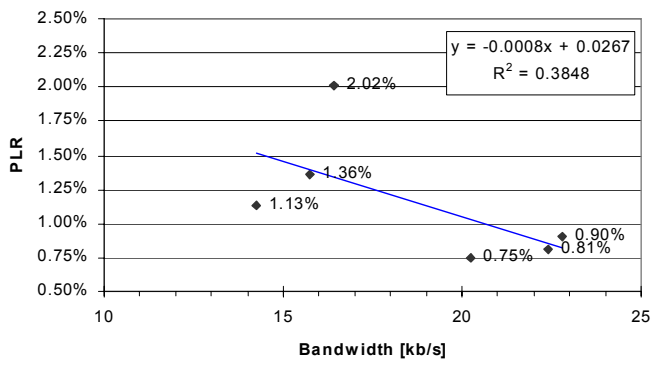


Figure 10. Packet loss rate versus bandwidth, networks A-F

VI. CONCLUSIONS

The paper describes a simple but powerful procedure for estimating key network and user oriented performance indicators. Estimating the quality of services for cellular data networks is a multidimensional process, requiring special attention while designing the test. Currently, acceptable performances must be customized based on technology. Measurements showed major differences in delivered QoS between technologies, deployments, as well as between fixed and mobile users. The IP layer performance (RTT, packet error rate) while giving a good estimate for the application layer bandwidth cannot substitute end-to-end QoS measurements.

REFERENCES

- [1] ITU-T recommendation E.800 (1994), Terms and definitions related to quality of service and network performance including dependability.
- [2] ITU-T recommendation G.1000 (2001), Communication quality of service: A framework and definition.
- [3] ITU-T Recommendation G.1010, End-user multimedia QoS categories
- [4] TS 23.107, 3GPP
- [5] J. Padhye et al., "Modeling TCP Throughput: A Simple Model and Its Empirical Validation," Proc. SIG-COMM'98, ACM, 1998
- [6] A. Papoulis, Probability, Random Variables, and Stochastic Processes, McGraw Hill
- [7] Alessandro Andreadis, Giovanni Giambene, Protocols for High-Efficiency Wireless Networks, Kluwer Academic Publishers, 2003