

Dyaptive SYSTEMS

DMTS-8000 Assisted Revenue Optimization



March 1, 2002

1	Introduction	1
2	Understanding Services: Traffic Models and Quality of Service Requirements	1
2.1	Circuit-Switched Voice	1
2.1.1	Quality of Experience	1
2.1.2	Grade of Service	1
2.1.3	Quality of Service	1
2.2	World Wide Web	1
2.2.1	Quality of Experience	1
2.2.2	Quality of Service	1
2.2.3	Grade of Service	1
2.3	File Transfer Protocol	1
2.3.1	Quality	1
2.4	E-Mail	1
2.4.1	Quality	1
2.5	Wireless Access Protocol (WAP)	1
2.5.1	Quality of Experience	1
2.5.2	Grade of Service	1
2.5.3	Quality of Service	1
2.6	Video Streaming	1
2.6.1	Quality of Experience	1
2.6.2	Grade of Service	1
2.6.3	Quality of Service	1
3	Defining and Testing Service Level Agreements	1
4	Differentiated Service Opportunities in CDMA2000 Networks	1
5	Revenue Modeling Framework	1



1 Introduction

The extent and rate at which the wireless services have penetrated the mass market is unprecedented. For a large segment of the population, cell-phones have become indispensable companions rather than novelties. As this market continues to attract more players, the competition is increasing and, consequently, offering network-differentiating services and controlling customer churn have become the main challenges for the service providers.

Taking a cue from the customer management experiences of Internet Service Providers (ISPs) and Telcos, the Wireless Service Providers have now started offering Service Level Agreements (SLAs) to exhibit their commitment to customer satisfaction.

Next generation services are being envisaged to capture the attention of consumers. Infrastructure and end-user devices are both being upgraded to deliver and deploy media rich content right into the palms “anytime, anywhere”, and to allow m-commerce and m-entertainment, push-to-talk, push-to-picture, etc. to mature and proliferate. The resilience and long term survival of a service provider in this highly competitive environment depends on the ability to capture this emerging wireless data market while retaining a solid base of voice subscribers. To a service provider, success through improved bottom line or ROI will therefore depend heavily on the adoption of these services at desirable profit margins.

Tough technical and financial decisions face 3G operators as they strive to capture these potential revenue opportunities that have grown from simple SMS type services to include wireless internet, web browsing, push-to-talk, ring-tones, Multimedia Message Service (MMS), video streaming, Location Based Services (LBS), etc. A financially considerate strategy is one that makes the most of the existing investments in widely deployed networks and manages the pace of future expansions of overlay coverage in accordance with the demand as well as the rate of acceptance of these services.

The network operator’s task is greatly simplified with tools that allow exploration of service and subscriber differentiation opportunities in their existing and planned deployments. These tools are needed to help quantify the resource requirements of new and existing services and the resulting implications on resource requirements and Quality of Service (QoS) and as a result, the implications for revenue and profit levels.

Further, SLAs are, in effect, an advertisement of network operators’ confidence in the performance of their networks. A network operator can reduce risk substantially with systems that automatically discover the performance bounds of the networks, taking into account vendor equipment differentiation. The operator can then define SLAs that take into account the tradeoffs between the value of guarantees with the cost of penalties and the cost of performance.

This whitepaper investigates the details of the above considerations and concerns, and proposes Dyaptive’s Mobile Terminal Simulator (DMTS) as a technology that not only tests a network for its technical merits but also verifies the sustainability of the business model of this next generation of wireless services. Based on the foundation of its unparalleled ability to load test a wireless network with synthetic but realistic traffic, the DMTS can be used to:

- map the conventional QoS metrics to the values that are more reflective of the true utility of the subscribers



- quantify the impact in terms of Return on Investment or profit gains for the service providers of
 - vendor equipment differentiation
 - new service offerings
 - service differentiation
 - resource management policies
- provide a platform for implementing and evaluating
 - potential service and subscriber differentiation opportunities
 - defining SLAs that can be confidently offered to the customers

The DMTS therefore allows service providers to realize the economic reality, in addition to technical feasibility, of their decisions before new networks or services are commissioned.

The white paper is organized as follows: first, the services that are currently being envisaged for wireless networks are studied; the traffic that these services inject into the network is characterized; and end-user's Quality of Experience (QoE) perceptions are quantified. Second, opportunities for service or subscriber differentiation are explored. The SLAs that could be potentially offered systems are outlined. The role that a DMTS plays in exploring service and subscriber differentiation opportunities or verifying that SLAs are technically possible, legally sound and economically viable is highlighted. Third, a DMTS-enabled econometric framework is proposed that evaluates various service offerings, QoS differentiation strategies, resource management policies, vendor equipment, and SLAs. This framework generates results in terms of net profit and revenue gain for the service provider taking into consideration the impact of customer churn as well as pricing. Finally some results are presented from tests conducted using the framework.



2 Understanding Services: Traffic Models and Quality of Service Requirements

A DMTS can generate accurate traffic streams for Circuit Switched (CS) as well as Packet Switched (PS) applications that are prevalent in wireless networks including voice, non real-time WWW, WAP/WML and FTP, near real-time video streaming, and real-time Voice over IP (VoIP) and Push-to-Talk (PTT). Some of the key traffic models possible with a DMTS are described below along with the traffic parameters that could be specified at run time to modulate the traffic characteristics. Also listed are the QoE parameters that are true representation of the operating quality of an application and are reflective of end user's experience of the service. These application level parameters are subsequently mapped to the network level GoS (Grade of Service) and QoS (Quality of Service) parameters that are ultimately controlled and fine-tuned to deliver the expected and desired QoE. The DMTS monitors all the application level as well as network level performance parameters listed below during the load testing of wireless networks.

2.1 Circuit-Switched Voice

Human conversation is a sequence of talk and listen modes. A conversational voice source thus follows the standard Markov source model defined in C.S0025 (TIA/EIA/IS-871). Voice codecs for spread spectrum networks adapt to this voice activity to reduce

the bit rate. Based on the underlying radio configurations, various voice call types should be generated including a combination of 2G and 3G radio configurations using the following traffic descriptors:

Traffic Descriptor	Distribution	Parameters	Default
call arrival	poisson	mean (calls per second)	
call duration	exponential	mean (seconds)	90
voice activity	exponential	mean	40%

Table 2-1 Circuit Switched Traffic Model

2.1.1 Quality of Experience

The ITU, in recommendations P.800 and P.830, defines subjective testing for "listening quality" and "conversational quality." Listening quality is a one-way phenomenon and is affected by the clearness, or clarity, and the loudness of the speaker's voice as it is perceived by a listener. The quality of two-way communication or conversation on the other hand is defined in terms of various subjective factors such as distortion, loudness, delay and echo. The network and the terminal both influence the subjective quality of voice. For example, distortion is the consequence of jitter buffer overflow and coding distortion at the terminal, and the bit error rate of the underlying channel. Similarly, the overall delay is not only the result of jitter buffer delay at the terminal but also the transmission and transcoding delay in the network. Echo also depends on the jitter introduced in the network as well as the efficiency of echo cancellation schemes in the terminal. ITU G.114 recommends a maximum one-way transmission time of 150 milliseconds for most voice applications whereas the acceptable levels of echo are prescribed in ITU G.131. In circuit switched systems, the delay and jitter incurred by a CS call is usually insignificant and, hence, the distortion, contributed by the loss in the channel and at the terminal, is the dominant factor in the perceptual voice quality.

The most widely recognized subjective quality assessment measure for conversational speech is the MOS (Mean Opinion Score) with listening quality scale (1 = bad to 5 = excellent). The performance of the system under test is rated directly as ACR (Absolute Category Rating) or relative to the subjective quality of a reference system as in DCR (Degradation Category Rating). The arithmetic mean of all the opinion scores collected under ACR is the MOS and under DCR is the DMOS (Degradation Mean Opinion Score). Objective quality assessment methods estimate subjective quality by measuring the physical characteristics of the terminals and networks. Several methods of objective quality assessment exist, each with a different aim, measurement procedure, inputs and the use of MOS. Notable among these include Opinion Model such as E-Model defined in G.107; speech layer objective models such as PSQM (Perceptual Speech Quality Measure), PESQ (Perceptual Evaluation of Speech Quality) and PAMS (Perceptual Analysis Measurement System); and packet layer objective models such as P.VTQ. The most comprehensive among these is the E-model that has 20 input parameters representing the terminal, network, and environmental quality parameters, and an output called the R-value. The R-value is an index of overall quality and has some correlation with the MOS.

Automated testing does not address Subjective Quality Assessment, however it can very effectively participate in the passive Objective Quality Assessment processes. The DMTS

does this by capturing the patterns of frame erasures of CS voice calls under the specified traffic load, traffic mix and channel conditions. These patterns are then used in E-Model to compute R-values.

2.1.2 Grade of Service

The usual parameters that quantify the GoS of CS voice services are:

- Call blocking rate
- Call dropping rate
- Call setup delay

Call blocking rate is the frequency that subscribers are denied access to the network and its resources. In non-spread spectrum systems, access is denied mostly when all the trunks or channels are already serving other calls and no spare capacity exists to handle an incoming call. In spread spectrum networks a call is blocked if either no spare or non-reserved (orthogonal) Walsh code is available at the BTS in the forward link, i.e. Code Blocking, or the call admission control scheme has computed that the interference caused by this new call will raise the Rise over Thermal (RoT) above some threshold and cause outage. It was a generally accepted fact that a BTS would run out of transmit power or receiver sensitivity before it would run out of Walsh space. However this was when the voice calls were still dominant in wireless networks. With high-rate data services continually on the rise, Code Blocking may become an unavoidable issue. Efficiency of code allocation algorithms in avoiding code fragmentation and, thus, preventing code blocking, under various anticipated arrival patterns of mix-mode traffic needs to be verified.

Further, under conditions of high call arrival rates, spread spectrum networks often fail to even recognize connection requests. This may happen when the connection requests on the reverse common channel are lost due to noise or collisions. Unfortunately, even though to a subscriber it is a clear case of access failure, for the network these events never happened, and as a result are frequently overlooked in many optimization efforts .

Call dropping rate is the frequency that ongoing calls get terminated prematurely. Calls are dropped if the network or the mobile experience excessive FER for a period of time. The network also terminates a call when the mobile is out of coverage for some time. This may happen due to cell breathing or simply because the mobile entered a coverage hole. Also, during handoff the likelihood of a call drop is increased. This happens when the sector or cell the mobile is migrating towards is unable to accept the handoff request. This can happen when the target sector or cell estimates that the additional call will cause a RoT that exceeds a threshold, when it has inadequate power to transmit the additional forward channel, or when it has insufficient Walsh codes to carry the call.

Call setup delay is the time lag between the instant when the mobile initiates a call or responds to a page and the instant when the application traffic starts flowing. The call setup delay depends on the traffic load and hence collisions on the reverse common channels and the queuing delay on the forward common channels, which in turn depend on the density of the mobiles in the System Zone or Location Area. The latency of the core network also contributes to this delay. The signalling channels need to be appropriately provisioned to keep the call setup delay bounded.

It is generally recommended that the network be provisioned to satisfy the call blocking probability of no more than 2% and even lower call dropping probability (1% or less), as



dropped calls are usually more detrimental to the end-user's perception of service quality. The call setup delay is recommended to be less than 10 seconds.

A DMTS tests a real spread spectrum network (in a lab environment) by generating realistic traffic load and channel conditions. During these load tests, the DMTS measures and reports the GoS as follows:

The DMTS counts how many originations were unsuccessful due to lack of resources and how many origination attempts or page responses did not even make it beyond the network access phase, thus helping network operators/planners to fine-tune coverage and capacity tradeoffs in their networks.

$$P_{blocking} = 1 - \frac{\text{CarriedLoad}}{\text{OfferedLoad}} = 1 - \frac{\sum \text{ServiceConnections}}{\sum \text{Originations} + \sum \text{PageResponses}}$$

A call that ends preceded by a "Release on Traffic" message from either of the parties is deemed normally terminated whereas any other pattern of termination is considered a dropped call. Histograms of different types of abnormal terminations are created to help network operators troubleshoot.

$$P_{dropping} = 1 - \frac{\sum \text{ReleasesOnTraffic}}{\sum \text{ChannelAssignments}}$$

Call setup delay is the elapsed time from the start of initial service request to the completion of any service negotiation in any of a mobile-to-land (ML), land-to-mobile (LM), or mobile-to-mobile (MM) scenario.

$$\Delta T_{CallSetup}^{ML} = T_{ServiceConnectCompletion} - T_{Origination}$$

$$\Delta T_{CallSetup}^{LM} = T_{ServiceConnectCompletion} - T_{PageResponse}$$

$$\Delta T_{CallSetup}^{MM} = T_{ServiceConnectCompletion}^{M2} - T_{Origination}^{M1}$$

The paging success rate and the rate at which calls are successfully directed to the voice mail are also measured if requested.

In addition to computing statistical mean and standard deviation of call setup delay observations during a load test, the DMTS can also determine the CDF (Cumulative Distribution Function) of the call setup delay or compute the probability:

$$P(\Delta T_{callSetup} \geq t) \text{ where } t, \text{ in seconds, is some relevant threshold.}$$

Since GoS is location dependent, the DMTS computes the above parameters on a per mobile basis; an aggregation over a sector or cell; or a normalization over a sector or cell cluster under consideration.

2.1.3 Quality of Service

As pointed out above, no other parameter has a more profound impact on the quality of CS calls than the FER. FER in turn is a consequence of the channel conditions as well as the SINR caused by other mobiles and sectors or cells that are transmitting at the



same time. For a network that is already planned and deployed, network operators can control SINR through Call Admission Control (CAC) and, to some extent, through Soft and Softer Handoffs.

One of the key capabilities of the DMTS is the ability to determine the pole capacity of CS Voice by determining the maximum number of voice users that a sector can support without causing system outage. System outage occurs when the number of simultaneous transmissions exceeds the capacity of a sector and therefore it is measured as the percentage of voice users in outage in either direction (i.e. reverse or forward links). The forward or reverse link outage per given voice user is evaluated using the short-term FER by measuring the FER over small windows (typically 400 ms). A single connection is considered to be in outage condition if the short term FER of a connection in the forward or reverse direction is greater than 15% for at least 1% of the connection time. If all connections show the aforementioned (per connection) outage condition at least 3% of the time, then the sector is considered to be in system outage.

The DMTS can accurately measure the FER as seen by the mobile on the forward link. On the reverse link, the DMTS does not have the direct visibility on the FER as seen by the BTS. However for Mobile-to-Mobile calls, since the endpoints are both the DMTS mobiles, any frame erasure on the reverse link of a mobile will result in a null frame on the forward link of the other mobile participating in the call. The DMTS can therefore monitor FER in both the reverse and forward links through its mobiles without requiring any performance monitoring on the network side.

2.2 World Wide Web

A typical web browsing session is divided into ON and OFF periods representing web-page downloads followed by reading times. The web-page downloads are referred to as packet calls, which represent a user's request for information. The reading time means the time required to digest the web-page before requesting another page in that session. A web-page is typically composed of several other objects, commonly referred to as embedded objects, such as images, audio clips and video clips. A web page download thus typically involves fetching the initial page (e.g. index.html) through an HTTP GET request, parsing this initial HTML page to find references to the embedded objects, and subsequent retrieval of these objects. A packet call itself has alternating ON and OFF periods. The retrieval of the initial page and each of the constituent objects occurs during ON periods whereas the parsing and processing times as well as the protocol overheads result in OFF periods. Even within this ON period of a packet call, packets do not arrive contiguously but intermittently due to delay and jitter resulting from scheduling and queuing in the radio access or core networks as well as the Internet.

HTTP uses TCP for transport and therefore the web requests and the web pages are transported as payloads of TCP packets. In HTTP/1.0, a separate TCP connection is used to transport each of the main and embedded objects of a web page. A configurable number of simultaneous TCP connections can be used for this download. In case the number of embedded objects in a page exceeds this threshold of the maximum allowable simultaneous connections then new TCP connections are initiated once the existing TCP connections close, to download these remaining objects. The flow control as well as congestion control overheads of TCP thus occur on a per object basis. In HTTP/1.1, on the other hand, a persistent TCP connection is used to download the initial page as well as the embedded objects of a web page that are located on the same server, and the objects are transferred serially over this connection. The TCP overhead thus occurs only



once per persistent connection. The DMTS supports HTTP 1.0 as well as HTTP 1.1 modes for web access.

Traffic Descriptor	Distribution	Parameters	Default
session arrival	poisson	mean (calls per second)	
session duration	exponential	mean (seconds)	
initial page size	truncated lognormal	mean standard deviation minimum maximum (bytes)	
embedded object size	truncated lognormal	mean standard deviation minimum maximum (bytes)	
embedded objects per page	truncated pareto geometric	mean maximum	
reading time	exponential	mean (seconds)	
parsing time	exponential	mean (seconds)	
number of web pages downloaded per session	geometric	mean (seconds)	

Table 2-2 WWW Traffic Model

2.2.1 Quality of Experience

The end-user's perception of the quality of a web browsing session over a network is shaped mostly by the time it takes to download a web page i.e. Page Download Time. This is the time elapsed between the instant an HTTP Get request is dispatched during a browsing session and the instant the last byte of the web page is received. Up to 15 sec delay is usually considered acceptable. The DMTS measures the page download time for each request from the web browsing sessions and reports the mean, standard deviation, CDF or probability $P(\Delta T_{Download} > t)$ where t , in seconds, is some relevant threshold.

2.2.2 Quality of Service

The QoS of a TCP based, non real-time packet services is measured in terms of throughput. There is always an expectation that some minimum throughput is maintained by every active TCP connection in the 3G sector. It is recommended that in a sector with multiple ongoing WWW sessions, no more than 2% of the users shall experience throughput less than 9.6 kbps [3GPP2 C.R1002-0]. The QoE requirement that the Page Download Time be bounded also puts an upper bound on the end-to-end TCP packet delay between the web server and the browser, which in turn depends on how well the

TCP slow-start and retransmission procedures respond to packet loss and timeouts. The performance of TCP thus depends on the choice of RLP scheme, Target FER for the data channels, data channel rate assignment and scheduling during the browsing session. Proportional throughput of TCP acknowledgements on the reverse channel also needs to be maintained so that the acknowledgements reach the transmitter without too much loss or delay in order to avoid activation of TCP congestion control.

The DMTS allows network operators to explore and evaluate RLP schemes, target FERs for data channels, data channel rate assignments and schedules and determine the optimum configuration given the expected background load and channel conditions. In addition, the DMTS also allows tuning of TCP parameters such as Maximum Segment Size, Initial Congestion Window, Maximum Window Size and use of SACK. The DMTS provides a comprehensive picture of the performance of the web browsing sessions by monitoring and reporting packet delay and loss at the TCP layer as well as the FER of the physical channels.

If both the source and the sink of data applications are controlled by the DMTS then packet delay is accurately measured. However, if the DMTS mobiles are to fetch the web pages from the sites located in the internet, the DMTS can report Round Trip Time as estimated by the underlying TCP protocol, instead of injecting its own probe messages in the network and thus avoiding unnecessary overhead. If available, the performance reports from network probes and third party Operations Systems can be correlated with the DMTS reports to improve the accuracy of QoS estimations.

2.2.3 Grade of Service

Typical of the “always-on” data services, web browsing sessions do not impose stringent call level GoS constraints on the network. A low probability of denying access is desirable when attempting to initiate data services (e.g., 2% or less).



2.3 File Transfer Protocol

An FTP session is a sequence of file transfers consisting of downloads and uploads, sometimes separated by reading times but mostly bulk transfers (e.g. mget or mput). This is also a reasonable model for other file transfer type applications such as e-mail, ring-tones and music download.

Traffic Descriptor	Distribution	Parameters	Default
session arrival	poisson	mean (calls per second)	
session duration	exponential	mean (second)	
file size	truncated lognormal	mean standard deviation minimum maximum (bytes)	
reading time	exponential	mean (seconds)	

Traffic Descriptor	Distribution	Parameters	Default
number of file transfers	truncated pareto geometric	mean maximum	

Table 2-3 FTP Traffic Model

FTP connection traffic is composed of signalling and data flows. FTP uses a reliable TCP connection for each of the signalling and the data flows. A new TCP connection is initiated for each file transfer. The signalling is used for connection setup, authentication, carrying file transfer requests to the server, and error messages to the client (e.g. File not Found).

2.3.1 Quality

Users expect the file transfers to complete in time that is proportionate to the file size. Smaller files, for example, should not take more than a few seconds; medium sized file transfers should complete within a few tens of seconds; and the larger file transfers should be no more than a couple of minutes. Because FTP is a TCP based application and the QoE of FTP (i.e. the File Transfer Time) imposes a delay bound, the GoS and QoS considerations of a file browsing session are similar to those discussed under Web Browsing. During a load test of a wireless network, the DMTS measures the performance of FTP sessions in terms of the following:

- QoE
 - File Transfer Time (seconds)
- QoS
 - Average throughput per subscriber per file transfer (kbps)
 - Packet loss (%)



2.4 E-Mail

An email browsing session typically starts with the email client connecting to the mail server and downloading the partial state of the user's mailbox (e.g. subject headings, senders and date-time of all the new messages in user's mail box). The user then downloads new email messages one by one, possibly interleaved with reading times. Occasionally, reading of an email is followed by a reply. Though less likely in case of wireless users, new emails may also get composed and sent during the session. On the average, the received emails are expected to be much larger than the emails composed and sent from handheld mobile devices. An email session is thus an initial download of the state of the user's mailbox followed by several downloads and uploads with read and sometimes compose times in between.

Traffic Descriptor	Distribution	Parameters	Default
session arrival	poisson	mean (calls per second)	
session duration	exponential	mean (second)	

Traffic Descriptor	Distribution	Parameters	Default
received message size	truncated lognormal	mean standard deviation minimum maximum (bytes)	
reading time	exponential	mean (second)	
composed message size	truncated lognormal	mean standard deviation minimum maximum (bytes)	
message composition time	exponential	mean (second)	
probability that a message is composed following a message read			
e-mail arrive	poisson	mean (e-mails per second)	

Table 2-4 E-Mail Traffic Model

A test can also be specified to take into account the relative impact of plain text, HTML or rich text on the email size, and the behavior of different e-mail clients and servers.

2.4.1 Quality

Again, the user expectation during the email browsing session is that the download or upload of emails (perhaps not the actual delivery of the email to the recipient) occur in matter of seconds. Furthermore, the emails should be delivered to recipients within 5 to 10 minutes. The % of emails not delivered within the aforementioned delay bound should be less than 5 to 10 %. Again, being a TCP based application with delay sensitive QoE, the GoS/QoS considerations are similar to HTTP.

2.5 Wireless Access Protocol (WAP)

The basic pattern of human interaction in a WAP/WML browsing session is no different from HTTP based web browsing i.e. page downloads interleaved with read times. A WAP gateway intercepts the request for a web page sent by a WAP/WML browser (deployed on the handheld device) using the WAP protocol; the request is translated into an HTTP Get request by the WAP gateway and sent to the destined web server; the main page and the embedded objects downloaded from the web server are formatted (resized, filtered and compressed) to meet the CPU/bandwidth/memory/display requirements of the wireless device and connection; and finally the re-formatted main page and the referenced embedded objects are returned to the WAP/WML browser using the WAP protocol. Each request from the WAP/WML browser thus causes the WAP server to send

back a response that is composed of a number of objects. Once all the objects associated with a page are received the reading time starts.

Traffic Descriptor	Distribution	Parameters	Default
session arrival	poisson	mean (calls per second)	
session duration	exponential	mean (second)	
object size	truncated lognormal	mean standard deviation minimum maximum (bytes)	
objects per page	truncated lognormal truncated pareto	mean standard deviation minimum maximum (bytes)	
WAP server response time	truncated pareto geometric	mean maximum	
inter-arrival time between objects	exponential	mean (second)	
reading time	exponential	mean (second)	
parsing time	exponential	mean (second)	
number of pages per session	geometric	mean	

Table 2-5 WAP Traffic Model

2.5.1 Quality of Experience

The QoE requirement of a WAP/WML based web browsing session is also that a web page download time is less than 15 sec.

2.5.2 Grade of Service

It is desirable that the acceptance rate of WAP/WML based web browsing requests is high. As in HTTP case, less than 2% probability of denying access should be maintained.

2.5.3 Quality of Service

The QoS of a WAP/WML based browsing session is measured in terms of throughput. It is recommended that in a sector with multiple ongoing WAP sessions, no more than 2% of the users shall experience throughput less than 4.8 kbps [3GPP2 C.R1002-0].

2.6 Video Streaming

A video stream is a concatenation of scenes of varying durations and activity levels. During a video streaming session the coded frames arrive at 1/fps where fps is the frame rate in frames per second. The frame rate need not stay constant during the session and may be varied according to the underlying channel and bandwidth conditions. The arriving frames are decoded and played out. A de-jitter buffer of a certain length is used in the mobile station to guarantee a continuous display of video streaming data. The de-jitter buffer allows for periods during which the real-time constraints need not be satisfied. The data is leaked out of this buffer at the source video data rate and filled as forward link traffic reaches the mobile station. The de-jitter buffer may run dry for some periods of time.

The frames are decomposed at the transmitter and transported over one or more packets that arrive at the receiver with inter-arrival delay introduced due to encoding as well as the queuing effects along the path. The frame size (bits per frame) varies from frame to frame and depends on the activity level or variability in the scene as well as the codec (e.g. H.264 or MPEG-IV) used. An encoded video stream therefore results in variable bit rate traffic.

Traffic Descriptor	Distribution	Parameters	Default
session arrival	poisson	mean (calls per second)	
clip duration	exponential shifted exponential	mean minimum (second)	
frames per second	deterministic	(frames per second)	
frame size	according to an auto-regressive process with a Gaussian excitation	AR process order mean standard deviation (bytes per frame)	
packet size	truncated pareto	mean maximum (bytes)	
packet inter-arrival duration	truncated pareto	mean maximum (bytes)	
number of P and B frames in a Group of Frames	deterministic		

Table 2-6 Video Streaming Traffic Model

2.6.1 Quality of Experience

Quality of experience from a video streaming session depends on various subjective factors including spatial resolution, temporal resolution, or color reproduction accuracy. MOS (Mean Opinion Score) is used to assess the subjective quality of a video sequence on a scale of 1 to 5. Degradations such as tiling, error blocks, smearing, jerkiness,

blurriness, object retention, jitter, frame-skip events, frame-freeze events are all included in the assessment. The Objective Quality Assessment methods include the mean squared difference of a pixel-to-pixel comparison between a degraded frame and its reference; the peak signal to noise ratio (PSNR) between a processed frame and its reference frame; a modified peak signal to noise ratio between a processed video frame and its matched reference; a peak signal to noise ratio computed using different weights for foreground and the background regions of frames; and a peak signal to noise ratio incorporating color. If the video streaming session also involves audio then the synchronization between the audio and the video streams also plays a significant role in the QoE.

As mentioned with voice, automated testing does not readily address Subjective Quality Assessment, but can effectively participate in a passive Objective Quality Assessment process by capturing the effect of network impairments on the video stream to estimate objective quality measures such as PSNR or MSE. Using a methodology similar to computing short term FER for voice, the DMTS also computes sequence wide short term distortion evaluation. For example, the entire video stream is partitioned into strips of 20 to 30 frames. An Objective Quality Assessment algorithm is applied to the frames. If the degraded frames constitute 15% of the strip and such strips constitute more than 10% of the entire stream then the stream is considered to have failed the perceptibility test. Mostly network impairments are considered and different jitter buffer sizes are emulated whereas other terminal impairments such as codec errors are ignored.

2.6.2 Grade of Service

Video streaming sessions do not impose strict call setup delay requirements. The call setup delay depends on the round trip time between the client and the server as well as the size of the de-jitter buffer. The access blocking should be maintained below 2% as for most of the call and session oriented services.

2.6.3 Quality of Service

The QoS of IP video streaming over a wireless network is measured in terms of packet loss, packet delay and jitter. Video streaming applications employ unreliable RTP over UDP to avoid the effects of congestion control and retransmissions of TCP but at the cost of incurring some packet loss. The packet delay and jitter is caused by the queuing delays along the IP path and the fluctuations in the bandwidth at the air interface due to data burst scheduling by the BTS. Unlike VoIP which is more sensitive to packet drop bursts than uniform packet drop during the call, the severity of the packet loss and the packet loss pattern during a video streaming session depends on whether the packet belonged to an Intra-coded (I), Predictive (P) or Bidirectional Predictive (B) frame and if any information related to motion compensation vectors is lost or not.

Using a DMTS, network operators can explore the benefits of using different RLP configurations or a lower FER target for the data channel to improve packet loss, and better data channel rates and schedules to improve delay and jitter, in a video streaming session given a set of traffic load and channel conditions. In addition, the DMTS also allows tuning of IP packet size to improve overall throughput.

The DMTS accurately monitors packet jitter. The accuracy of the end-to-delay and packet loss measurements during the load test depend on whether the video streaming sources and sinks are both controlled by DMTS. In case the mobiles are instructed to download video streams from the media servers located in the internet, then the Real



Time Control Protocol (RTCP) reports can be used to estimate the performance of the RTP layer. Further it is possible to inject probe packets to detect any performance bottlenecks along the path; or, if available, to correlate reports with the performance reports from network monitoring probes and OSs deployed in the network to create an accurate picture of one way end-to-end delay and packet loss.

3 Defining and Testing Service Level Agreements

A Service Level Agreement (SLA) is a contract between either two service providers or a service provider and a customer that specifies the QoS level that can be expected and the penalties the service provider will pay in case the QoS commitment is not met. The key steps in an SLA management lifecycle include:

- Service Level Specification
 - It is the process of developing the definition and specification of parameters that describe the service. These parameters include availability, reliability, latency, and loss.
- SLA measurement
 - The QoS that the service providers promised and delivered to their customers needs to be measured accurately.
- SLA compliance reporting
 - The subscribers need to be notified that they are receiving the service levels that were committed to them.
- QoS management and control
 - The network must be provisioned and appropriate controls must be in place to ensure that the established SLAs are honoured.

Before analyzing the feasibility of SLAs for wireless networks, let's revisit some experiences with SLAs for wired networks. The telephone operating companies are able to provide very tight performance assurances for circuit switched digital data services. Among the packet switched networks, ATM and frame relay are designed to offer relatively robust performance assurances. This is because these networks employ call admission control and flow and congestion control mechanisms. In addition, these networks, at the time of connection establishment, expect each connection to specify the profile of the source traffic, and employ traffic shaping and policing mechanisms on ingress traffic at the User Network Interface (UNI) as well as the Network Network Interface (NNIs). An SLA in these networks therefore includes the following:

- traffic descriptors that describe the traffic envelope in the form of (σ, ρ) , constraints or a token bucket
- specification of the treatment of the packets such as dropping, shaping and re-marking for packets that do not conform to the agreed upon traffic envelope,
- the QoS guarantees offered by the network to the connection e.g. delay, jitter, packet loss and throughput guarantees, and
- the contract schedule i.e. hours of the day, month, and year when the contract is binding or applicable.



IP networks, on the other hand, operate on a “best effort” basis and can only offer loose assurances. The SLAs in IP networks use performance metrics that are based on large time scales (one month is typical) and are usually offered with the condition that the average utilization of the access links be less than 50% or so. This is because the access links are relatively low bandwidth and are the likely congestion points because of the connectionless routing and lack of robust call admission control in IP networks. More concrete SLAs are usually offered only for the IP backbone where Frame Relay, ATM or circuit-switched SONET are used as the underlying transmission layer.

The next evolution of SLAs in IP networks will be achieved when the IntServ and DiffServ protocols become widely deployed. These technical foundations will enable inelastic applications such as VoIP, that require tight QoS guarantees, to coexist with non real-time TCP flows in IP networks and allow absolute or statistical performance bounds to be offered over smaller timescales e.g. “95% of all packets measured over observation windows of 5 minutes will cross the domain in less than 10 milliseconds”.

For an end-user the parameters of interest are loss, delay, and jitter whereas an enterprise customer usually seeks bandwidth and availability (Mean Time Between Failures, Mean Time To Recover) commitment from the service providers. A typical SLA of network availability may be worded as “The network will be available 99.99% of the time over a period of 1 year”. In addition, performance guarantees of QoE, GoS or QoS are also offered to support business critical applications. Some specific examples include:

- Response time – “95 percent of users will experience a response time of less than 12 s during normal work hours i.e. between 8 a.m. and 6 p.m.”
- Throughput – “No more than 2% users will receive a throughput of less than 9.6 kbps during working hours”.
- Call blocks – “call blocks are less than 2% measured over a period of 1 month”
- Call drops – “call drops are less than 1% measured over a period of 1 month”
- Packet loss Rate – For real-time services such as VoIP, the Packet Loss Rate guarantee could be offered as “Less than 0.1% for a 5 min sample” whereas for non real-time services a possible guarantee may look like “Less than 0.1% for 95% of the collected samples”.
- Packet delay – For real-time services Packet delay guarantee is expressed as “Less than 150 ms for a 5 min sample”, but for non real-time services, again, these could be worded as “less than \leq 50 ms for each 95% of the collected sample”.
- Packet Jitter – Packet Jitter is not applicable to non real-time services whereas for real-time VoIP services a guarantee of “less than or equal to 45 ms” is expected.
- Utilization — “35 simultaneous users will be supported during peak hours”.

The absolute QoS guarantees can only be offered if the stimulus to the network (i.e. the traffic) as well as the network’s response to it is predictable. As a rule of thumb, if the traffic has a deterministic traffic profile (description of its temporal properties) e.g. (σ, ρ) constraints, then absolute performance bounds could be guaranteed. Contrarily, if the traffic can only be defined in terms of statistical averages (e.g. mean and standard deviation of bit rate etc) then QoS also can only be guaranteed in terms of statistical



bounds. In IP networks, for example, the traffic load is always unpredictable due to lack of effective traffic control mechanisms. Also, since IP networks are connectionless, the packets from the same session can take different routes, and therefore the amount of traffic on any given link at any given instant cannot be predicted. As a result QoS guarantees in the current generation of IP networks cannot be offered.

It is apparent that cellular networks cannot possibly be expected to offer QoS guarantees as well. This is because unlike their wired counterparts, the cellular networks suffer from inherent unpredictability of the underlying wireless channels. Coverage holes caused by line of sight (mobile in an elevator etc) or cell breathing deteriorate service quality and render availability unpredictable. Even though 3G networks are armed with effective power control as well as admission control mechanisms to keep the influx of traffic in check, the variations in the wireless channel characteristics unfortunately can neither be controlled nor predicted and, therefore, the absolute performance bounds cannot be guaranteed. The capacity of a CDMA sector, as opposed to its TDMA/FDMA counterparts, is also soft, and, at any given time, depends on the interference from the traffic within the sector as well as in the neighboring sectors. The QoS guarantees therefore can only be offered in terms of statistical performance bounds accompanied by various traffic and environment conditions that must prevail for these performance bounds to materialize. Furthermore, in case violations of SLAs take place, the service providers must be able to prove that the violations didn't occur because of inadequacies of their resource management policies but due to unknown environmental factors beyond the scope of their resource management system. These SLAs are likely to include conditions such as "RoT (Rise-above-Thermal) is less than 5db" or "the average utilization of the sectors in the area is less than 50%". To take the spatial dimension of cellular networks into account, a topological scope should also be included in the SLAs to specify the geographical area where the SLA is applicable.

Another major challenge for cellular networks is to support end-to-end SLAs. The wireless service providers almost always require services of other service providers such as leased lines and trunks to transport backhaul traffic, and almost always a session through a wireless link also traverses either the Public Switched Telephone Network (PSTN) or the Internet. This includes the effect of roaming which is often hidden in many "national" networks. The onus is on wireless service providers to propagate the conditions of end-to-end SLAs across multiple vendors, multiple domains and multiple providers.

A DMTS is an indispensable component of SLA management infrastructure for wireless networks. In case SLAs have already been drafted, a DMTS can automatically do the due diligence to determine if the network deployment has the necessary resources to allow SLA-stipulated aggregate traffic to be carried at the satisfactory level of QoS. Alternatively, a DMTS is an effective platform to drive a trial network to its performance limits and derive the SLAs that could be offered confidently.

An SLA can be drafted from the service provider's perspectives or from a subscriber's perspective. While the large time scales protect the service provider, the end user is not offered a true level of assurance. During peak hours (i.e., the busy hours between 9:00AM and 5:00PM), it is possible that a sector is heavily congested with resulting in a high likelihood of performance degradation. However if the latency and utilization metrics are averaged over large time scales, the SLA might not have been technically violated.

Similarly, in order for the 98% availability guarantee to be violated, the network has to be physically down 174 hours during a one-year period. Besides, depending upon the



wording in the SLA, it is possible for the service providers to avoid penalties in cases where performance deteriorates rendering a network unusable though not actually physically down. Subscriber oriented availability metrics, on the other hand, will assume that the network is unavailable once the network becomes unusable. Using a DMTS, optimal time scales can be derived for SLAs that ensure subscriber satisfaction while protecting service providers against penalties.

A DMTS also helps determine if application level performance parameters e.g. QoE metrics should be used in SLAs or whether GoS/QoS parameters are more appropriate. An application level performance metric offers several advantages. Firstly, an application level metric, such as Page Download Time defined above, implicitly incorporates GoS and QoS parameters into the performance assessment. Secondly, an application level metric provides a natural assessment of network's performance from an end-user's QoE perspectives as compared to conventional QoS metrics such as packet loss or latency. Using a DMTS, a network deployment could be loaded with anticipated traffic load and traffic mix, and the impact of the network's performance on the end user's QoE can be measured and appropriate SLAs can be subsequently discovered. Some examples of such discovered SLAs are "95% of the Page Downloads will be less than 2 seconds over a two hour time period provided the average RoT remains below 5dB" or "the maximum number of simultaneous web users is 30 if on the average the cell and its 1st tier neighbors are loaded less than 50% of their soft capacity".

4 Differentiated Service Opportunities

Support for differentiated services in wireless networks is essential if the next generation of services is to coexist with current offerings in the critically limited wireless spectrum. Most of the deployments however offer minimal explicit support for service differentiation. Careful study of the relevant standards though reveals that several protocols and parameters can be manipulated to recognize different categories of subscribers and realize service differentiation without incurring any additional overhead. A DMTS proactively explores such possible means of subscriber and service differentiation and verifies that selected groups of subscribers or classes of services continue to receive preferential treatment as far as allocation of resources in concerned even under heavy load or congested conditions. This section highlights the underpinning of subscriber, service and traffic flow differentiation and the role the DMTS plays in its utilization.

Typically, access to the network can be controlled and differentiated if certain Medium Access Control (MAC) parameters are assigned different values for different classes of subscribers or services. These parameters include back-off window size, persistence testing, as well as other random access parameters. A group of mobiles using a larger back-off window, on the average, will suffer longer delays subsequent to a collision on the reverse, random access channel resulting in an overall longer call setup delay. If power controlled access is available, the mobiles that are assigned higher power, on the average, will also access the network earlier than the mobiles that are kept at a lower power due to capture effect during collisions. With support for multiple forward common channels, the call setup delay for mobile terminated calls could be further differentiated by properly configuring the channel assignments to the mobiles. In other words, certain channels could be reserved for high priority subscribers so that the queuing delay for the page messages destined to these subscribers, and hence the overall call setup delay, is reduced.



Once a request for service connection is successfully accepted by the network, the network assigns resources based either on the subscriber identity, the service type or any explicit resource requirement specified in the request. The network can differentiate throughput, delay and loss requirements across service types, or even among the connections of a particular class of service, by controlling the allocation of Walsh codes, power, times, and durations of packet channels, and retransmission schemes. For example, upon receiving the data for both the high priority as well as the low priority connections at the same time, the network can allocate high rate channels and for longer durations for delay sensitive applications whereas buffering the data for best effort services for later delivery to keep the interference to the minimum. Similarly, the power assignments to the forward link channels can be assigned in proportion to the target FER of the Class of Service (CoS). Finally, a more aggressive retransmission scheme could be used for services with stringent loss requirements.

With a DMTS, the service providers and equipment vendors alike, get a flexible platform to explore and verify the aforementioned means of realizing service differentiation in their systems. The notable feature is the flexibility with which the DMTS allows end-users to manipulate its software-defined test terminals (STTs) to create real-world scenarios with multiple classes of services and categories of subscribers accessing the wireless network with distinct configurations, traffic descriptors and QoS requirements; and its ability to accurately measure the QoE, GoS, and QoS received by each class of service or subscriber. As an example, different groups of STTs can be instantiated by a DMTS, with each group configured to use a distinct back-off window size to access the network. The relative gain in the call setup delay achieved by STTs employing smaller back-off windows at the expense of STTs employing larger back-off window is then reported for further consideration by network operators. Other reverse channel related parameters, identified above, can similarly be explored for service differentiation individually or collectively.

Further investigation into the relationship between CoS and FER are possible. Instantiated STTs can be grouped and configured to request a target FER according to their CoS. The DMTS then automatically determines the spectrum of traffic and channel conditions under which the requested target FERs are exactly achieved, and the spectrum of conditions in which the requested FERs are not achieved but the STTs continue to receive preferential treatment in the form of better power allocations than others.

Finally, DMTS comprehensively verifies if the Call Admission Control (CAC) or Resource Allocation policies of a network offer any service differentiation. Most networks limit the maximum bit rate that a subscriber can receive. For example some subscribers could be restricted to no more than a basic channel rate whereas preferred users could be authorized to receive a higher rate provided the conditions are favorable. Again, a DMTS not only conducts conformance testing of such restrictions but also discovers the pattern with which the bit-rate of various classes of subscribers deteriorates as the network becomes congested. In other words it verifies whether the network schedules high rate channels in a fare-share manner; further, if the result is positive, then the DMTS also shows the impact of the resource allocation scheme on the packet delay of the overlying application session. The results of such tests provide useful information and insight when drafting SLAs.

Traffic from popular data applications such as FTP, HTTP, WAP, video streaming and VoIP is composed of not only one or more user data flows but also at least one



application level signalling flow. Traffic belonging to a VoIP session, for example, includes a SIP flow which carries application layer signalling and an RTP flow which carries voice data. Both flows have distinct QoS requirements. Unfortunately, many networks recognize flows only at the granularity of a subscriber or mobile. Once a mobile establishes a connection with the network and a data tunnel is created between the mobile and the PDSN or SGSN, all the flows emanating from all the applications that are active on the mobile are multiplexed and transported over the single tunnel. The target FER, channel rates and retransmission schemes are therefore negotiated for the underlying physical channel rather than individual application level flows. It may be possible to map the QoS requirements of individual applications or individual flows within each session of the application to a target FER, retransmission scheme or set of channel parameters. Accordingly, the PDSN or SGSN, though it has clear visibility to these flows, makes no attempt to differentiate applications or flows within an application to offer any preferential treatment. The DMTS can be used by application developers and equipment vendors to explore application provisioning in which differentiated QoS could be offered to different flows within the application. Some considerations to this effect include:

- Allocating more power to traffic channels immediately following a data connection establishment to ensure that the tunnel connection setup and Mobile IP registrations occur without any loss and the user is able to connect to the Internet within the acceptable delay.
- Sending application level signalling messages over a circuit channel, as opposed to a packet channel. This is because the circuit channel typically has lower target FER and hence more power allocation. SIP messages of a VoIP session should be sent over the circuit channel and the voice data over a packet or circuit to realize improved call setup performance for services like Push-to-Talk.
- Using an aggressive retransmission scheme during connection setup and thereafter going back to a default one once the connection is established. For example, requesting an aggressive scheme during the three-way handshake of TCP connection setup so that SYN packets are not lost and have to be retransmitted after a timeout, and then reverting back to normal scheme once the ACK packet is received and the connection is established.
- Using data bursts over common channels for transmitting vital data while the underlying traffic channel is transitioning to an active state. Applications like Push-to-Talk, whose token requests to invite others (via SIP messages) could be sent via such bursts while the traffic channels are transitioning to an active state, will gain considerably in terms of latency.



5 Revenue Modeling Framework

Load testing in the past has mostly focused on assessing the technical merits of the network. A network, set up in a lab, is loaded with synthesized traffic of anticipated intensity and mix, and its response is measured in terms of various QoE, GoS, QoS parameters. One of the motivations behind this is to verify that the network is provisioned to deliver the expected quality of service to the end user and help maximize its utility to the subscribers. Such customer-centric objectives however fail in two major ways. First, they give no clear guidance on how the service provider should trade off the performance in one area (e.g. sector throughput) for another (e.g. call blocking rate). Second, they fail to capture the network's utility to the service providers, which is the Return on Investment or the net-profit. To a service provider, a healthy business model through which new

services can be profitably introduced is absolutely necessary and, therefore, testing a network for revenue optimization is of utmost significance and relevance.

For example, traditionally the performance of a network is measured in terms of overall throughput. While in a monolithic, voice-only network that does not support price-based service differentiation, throughput is directly proportional to revenue. Multi-service 3G deployments with non-uniform traffic, channel conditions and service coverage may not always support this proportionality. In these networks with economically efficient resource allocation schemes in place, objectives such as maximization of throughput may not necessarily result in revenue maximization and, instead, relative yield of each service needs to be taken into consideration. The revenue that services generate for service providers depends on their usage. The more subscribers per service a network can accommodate, the higher will be the revenue for the service provider. In interference-limited 3G networks though, eventually customers are will be denied access to network resources to alleviate onset of congestion, or else outage conditions may prevail impacting all connections. Either situation would render dissatisfied customers some of whom may decide to switch to new service providers (customer churn). Pricing and differentiated QoS are used to distinguish and control demand and thus maximize the network utility for the service provider.

Measuring a network for revenue optimization is anything but a trivial extension of the functionality of conventional load-testing solutions. In fact, verifying that a multi-service network has a sustainable business model and that the network is configured to yield maximum ROI requires development of an extensive framework that allows modeling of numerous socio-economic and technical aspects as well as their inter-relationships. To provide true value to the service provider, such a framework must take into account the effect of pricing on the usage of services, support for differentiated services and its impact on customer satisfaction and churn, and the effect of new services on the operational costs to name a few. Some of these factors are not easy to model. Let's take for example the effect of pricing on service demand or usage. The amount of traffic for a service that customers are willing to pay for defines the demand for that service. Intuitively speaking, demand for a service is a decreasing function of price. When the service is cheap its usage is high but if the service is expensive then it's likely to generate less demand. At these two extreme ends of the price-demand curve there is no price-based control of usage. The price-demand curve also depends on the economic-demographics as wealthy users are usually willing to pay more for the same usage. Furthermore, charging could be flat rate, dynamic (congestion dependent) or prepaid, and prices may have temporal component (different prices for different time of day); and each of these aforementioned pricing criteria impact the service-demand in a unique way. The important thing is to realize that somewhere between the two extremes of price-demand curve, must exist a region where the demand or service-acceptance does vary with price across all demographics and pricing schemes. It is quite plausible that a precise relationship between price and demand for the new service offerings may never be known in advance. Nevertheless, typical of any economic or financial decision making, service providers are expected to have assumed the existence of such relationships and are likely to have derived valuable insights to model such relationships based on historical trends and observations.

The DMTS provides a comprehensive off-line network testing and engineering framework that effectively reconciles the financial as well as the technical merits of a 3G deployment. Any new service or SLA offering or consideration for equipment, software or protocol replacement or upgrade, or subscriber / service differentiation can be measured for



revenue optimization using a real 3G deployment and a DMTS. The aforementioned framework allows testing and evaluation of a network in terms of the Net Gain for the service providers. The net gain is a function of Revenue, Operational costs, Churn cost and SLA Penalties, and is computed as:

$$Gain_{SP} = R - (Cost_{Op} + Cost_{Churn} + Cost_{SLA})$$

where:

R is the revenue from the network

Cost_{op} is the operational costs required to run the network

Cost_{churn} is the cost resulting from the churn of customers

Cost_{SLA} is the cost resulting from penalties due to SLA violations

A network engineer or service planner aiming to explore various pricing, demand, network design, traffic engineering and service provisioning options can use a DMTS to determine the real value proposition that these options will render to the service provider instead of getting tangled into the Cost, Coverage and QoS web or being bogged down with a nightmarish task of trying to grasp, and make sense of, almost infinite number of performance measures and their interdependencies.

Given a set of CoSs and the pricing and demand models associated with each, Given the offered service types and either subscription per service per unit coverage area and the price-demand relationship, the DMTS creates software-defined test terminals (STTs) that are grouped into specified Classes of Services (CoSs) or categories of subscribers. It then generates traffic of anticipated temporal variations; executes simulation runs iteratively; and estimates the net gain the configured scenario will render to the service provider if reproduced in the field over a period of time.



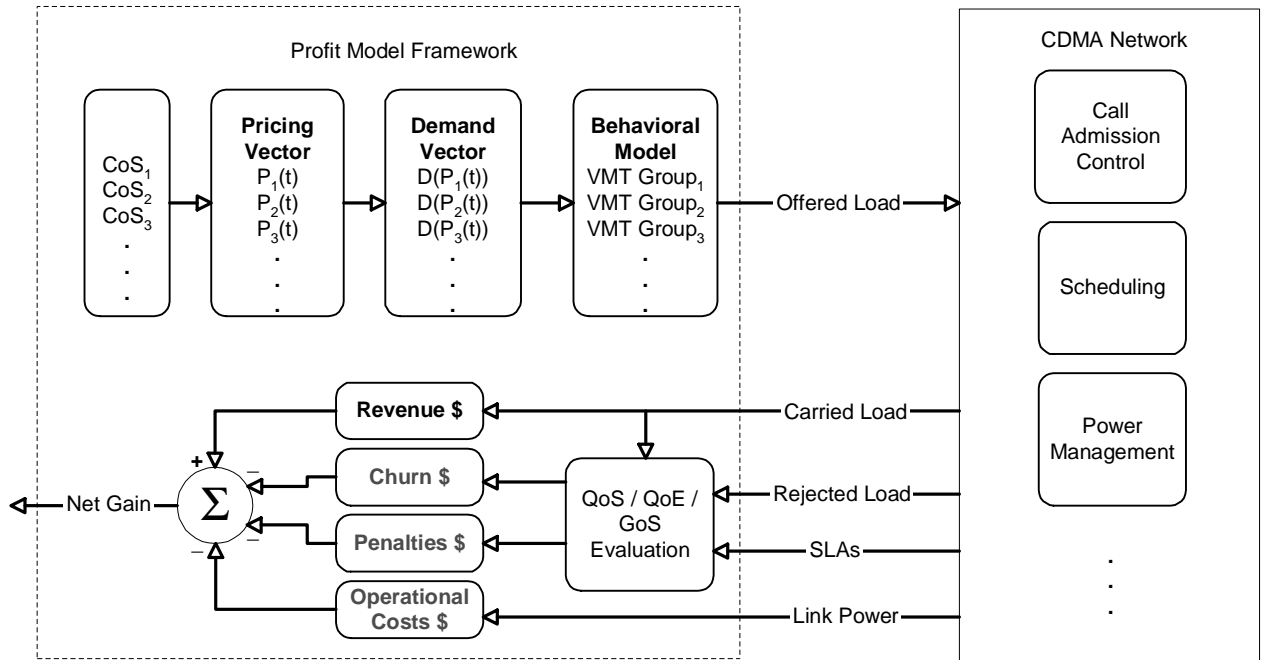


Figure 5-1 Profit Model Framework

As mentioned earlier, although attempts have been made to formulate the price-demand relationship and some deterministic models have been proposed by economists, service providers are expected to have quantified the impact of pricing on the demand through historical trends, customer interviews, or focus groups. In case such information is not available, the projected traffic load per service per unit coverage area could simply be specified to the DMTS. The revenue and cost are then computed as follows.

Let $n_{jk}(t)$ represent the number of calls or sessions of CoS j that are in progress at time t in sector k being driven by the DMTS, and r_j is the revenue in dollars per second generated by a call of CoS j . The Service Provider then earns overall Revenue at a rate

$$R(t) = \sum_k \sum_j n_{kj}(t) \cdot r_j$$

From a network engineering perspective, the operational cost of interest is the power expenditure. DMTS monitors the total power on the forward link. Let p_k be the rate at which the power is being expended by the network to support all the traffic and signaling channels in the forward link of a sector k . The rate of expenditure in dollars per second is then

$$Cost_{Op}(t) = \sum_k p_k(t)$$

Churn is the rate at which dissatisfied subscribers leave a service provider. Again, churn is also a consequence of numerous psychological, sociological and economical factors (e.g. terms of subscription contracts etc) which are difficult to characterize. For this

present purpose, the objective is to establish a legitimate and measurable relationship between the service quality and churn. For a wireless subscriber, the decision to churn can be perceived as a result of an imbalance between the utility that the subscriber receives from a service and the price paid for the service. Expressed another way,

$$\text{Churn} = f(\text{Utility} - \text{Price}).$$

Thus, either deterioration in the subscriber's utility or an increase in the service price could cause a subscriber to churn. Utility is a measure of satisfaction that a subscriber receives from the service and is a function of service quality. The utility can be measured in monetary units so that it could be compared to the price when estimating churn.

Specifically,

$$U = K \cdot f(\text{Service Quality})$$

where K is a constant of proportionality and Service Quality is either measured in terms of application level QoE parameters (e.g. Page Download Time) or in terms of network level GoS or QoS parameters e.g. call or session blocking rate, packet loss, packet delay and jitter .

For delay-sensitive VoIP, PTT or even MobileVirtual Private Network connections, utility is a decreasing function of delay, so that more delay means less utility. For non real-time data applications such as WWW, FTP and email, utility increases with throughput.

A mapping of QoE, GoS, QoS parameters to subscriber utility can be specified as part of a DMTS test. For example, utility may be a linear, exponential, lognormal or a sigmoid function of one or more QoE, GoS, or QoS parameters, or the relationship between the utility and service quality may be defined as a set of "If-Then-Else" rules or a simple table. Further, a DMTS has the ability to support live application sessions with a synthetically generated background traffic load. This capability can be used to estimate and establish such utility models empirically.

Intuitively speaking, there is always a limit beyond which any further improvements in QoS do not result in any gain in utility. A simple mapping then perhaps could be as follows:

- For a CS Voice Subscriber belonging to CoS j , the utility $U_j^{CSVoice} = 0$ if the call is blocked, the call setup delay is unacceptable, the call is dropped, or the short term FER of the call is unacceptable, otherwise it is r_j .
- For a WWW browsing session belonging to CoS j , the service utility $U_j^{WWW} = 0$ if either the Page Download time or the received throughput during a Packet Call (Page Download) are unacceptable. For FTP, Email and WAP/WML services similar definitions of utility might be used.
- For a video steaming session belonging to CoS j , the utility $U_j^{Video} = 0$ if any of the Objective Quality Assessment criterion identified above fails, otherwise it is r_j .

In the above assumed utility model either the subscriber receives full utility or no utility at all. The undercurrent here is that if a subscriber requests a service, it is reasonable to assume price at which the service is being offered is acceptable. Thus the subscriber could be assumed to have recovered the investment of r_j if the desired QoS is delivered. If on the other hand the received QoS didn't meet expectations, the investment is deemed lost. The utility in this case therefore is a gated function of QoS.



Let u_{jk} be the % of CoS j connections for which utility is 0 at any instance during a DMTS assisted load test of sector k . Since QoE, GoS, and QoS are location specific, so should be the utility and, therefore, if the load distribution among these sectors is non-uniform then a weighted average is computed (to give more weight to the congested sectors/cells) i.e.

$$\hat{u}_j = \frac{\sum_k \lambda_k \cdot u_{kj}}{\sum_k \lambda_k}$$

where λ_k is the aggregate call arrival rate in sector/cell k . The term call is used for voice calls as well as packet calls. The term \hat{u}_j therefore represents the fraction of potentially dissatisfied customers in the network and can be estimated based on instantaneous, short-term or long-term measures of utility during the simulation runs. For short-term measurements, the utility is based on a small window of consecutive connection attempts whereas long-term measurements are taken over the entire simulation run. The churn rate is then proportional to \hat{u}_j . One possible quantification of lost revenue due to churn could be then

$$Cost_{Churn}(t) = \sum_j \frac{\lambda_j}{\mu_j} \cdot Ch(\hat{u}_j) \cdot r_j$$

where λ_j is the mean call arrival rate of CoS j calls in a sector/cell and μ_j is the average duration, $Ch(\hat{u}_j)$ is the churn rate, and r_j is the average rate of revenue lost due to a missed CoS j call. For real-time, ON/OFF packet calls with tight bounds on packet delay, the average revenue per second is computed based on the equivalent bandwidth concept. Data calls involving elastic traffic generate revenue based on bytes transferred per observation interval.

Similarly, a DMTS test measures the durations during which SLAs are violated during the load test and includes the penalty costs in the revenue model. The loss due to SLA violations is an immediate financial loss whereas the loss due to churn is the anticipated future loss. For a simulation run of duration T , the average rate of net gain is

$$\bar{Gain}_{SP}(t) = \frac{1}{T} \int_0^T Gain_{SP}(t) \cdot dt$$

The above figure could then be used to project net gain for the service provider over longer periods of time.

Any prospective network design, or any alterations to an existing one, could be pre-verified to improve the net gain for the service provider using such DMTS-based tests. The revenue model proposed here is parametric and can be customized to accommodate other billing criterion as well. For example, if an “unlimited calls” based billing regime is assumed, then even though the rate of revenue for the service provider will be constant (proportional to the subscription size), the emphasis in these situations will be to maintain low resource consumption while keeping churn and SLA violations in check to improve the net gain for the service provider. As one can see, often these requirements will be in conflict, but the DMTS-based test provides the ability to weigh these choices on a single scale and thus making it easier for the network operator to make planning, service rollout and service differentiation decisions.



If the network is lightly loaded then financial rewards of accepting new calls clearly outweigh any associated operational costs. However, if the network is operating under congested conditions then the benefit of revenue received due to this new call must be weighed against its side effects. Any new call will produce interference to other calls in the same as well as neighboring sectors. This interference may result in deterioration in QoS causing increased churn, or even SLA violations. If the network decides to accept the call while keeping acceptable QoS, the BTS will need to allocate higher power to connections causing exponential increase in operational costs. Furthermore, the increase in the power at the forward link of a BTS will also cause interference to the neighboring sectors and may force these BTSs to increase their power expenditure as well. The DMTS automatically measures and tracks such dynamics during a simulation run and explores if the call admission control schemes of a 3G network are able to produce the maximum possible gain to the service provider under projected busy hour traffic load conditions.

In case the network operator contemplates increasing the number of soft-handoffs to improve capacity in a target coverage area, the DMTS can automatically determine that the network design and traffic load allow for any improvement in the net gain for the service provider. This is because even though soft-handoffs reduce interference in the reverse link, all the BTSs involved in the soft-handoff need to maintain traffic channels in the forward link for this call causing an increase in the network wide power and Walsh code consumption.

The DMTS-based test also assists in network provisioning decisions such as increasing the number of signaling channels at the expense of the traffic channel quota in order to improve call setup delay, or assigning lower target FER to some CoSs at the expense of others to achieve better call quality. The overall operational costs such as power and Walsh space consumption may stay the same but the DMTS can effectively find out any drop in revenue is compensated through improvements in churn and SLA violations. Similarly other network design, optimization, service provisioning and service differentiation considerations can be evaluated using this DMTS-enabled framework.

A key concern for service providers as they venture into new service offerings is the uncertainty in demand. The network engineers have to decide on a resource allocation strategy that strikes a balance between services such as voice that may be lower-priced but have an established subscriber base, and the new offerings that may be priced higher but have no guaranteed traffic volume. The DMTS helps assess this risk by automatically and iteratively generating traffic loads covering the entire probability distribution of the demand and measuring the expected revenue as well as the risk of revenue shortfall in a 3G deployment provisioned for differentiated services. The network operators can then mitigate the risks of revenue shortfall by fine tuning the resource allocation to maximize a weighted average of expected revenue and revenue shortfall.

It may also be worth mentioning finally that a considerable portion of billable minutes are estimated to never even make it to the billing systems or, if they do, are billed incorrectly. When a DMTS loads a 3G deployment and monitors its performance, it also drives other third party Network Management Operations Systems, including billing systems. The DMTS can prevent such revenue losses by facilitating the monitoring of the loss of Call Data Records before such systems are deployed in the live network. The DMTS and a profit-based measurement framework therefore are effective tools of end-to-end Revenue Optimization for the service providers.

