Wireless Netw (2008) 14:233–246
DOI 10.1007/s11276-006-9232-0

Performance modeling and evaluation of data/voice services in wireless networks

Jae-Hyun Kim · Hyun-Jin Lee · Sung-Min Oh · Sung-Hyun Cho

Published online: 9 October 2006

Abstract Application-level performance is a key to the adoption and success of the CDMA 2000. To predict this performance in advance, a detailed end-to-end simulation model of a CDMA network is built to include application traffic characteristics, network architecture, network element details using the proposed simulation methodology. We assess the user-perceived application performance when a RAN and a CN adopt different transport architectures such as ATM and IP. To evaluate the user-perceived quality of voice service, we compare the end-to-end packet delay for different vocoder schemes such as G.711, G.726 (PCM), G.726 (ADPCM), and vocoder bypass scheme. By the simulation results, the vocoder bypass scenario shows 30% performance improvement over the others. We also compare the quality of voice service with and without DPS scheduling scheme. We know that DPS scheme keep the voice delay bound even if the service traffic is high. For data packet performance, HTTP v.1.1 shows better performance than that of HTTP v.1.0 due to the pipelining and TCP persistent connection. We may conclude that IP transport technology is better solution for higher FER environment since the packet overhead of IP is smaller than that of ATM for web browsing data traffic, while it shows opposite effect to the small size voice packet in RAN architecture. We show that the 3G-1X EV-DO system gives much better packet delay performance than 3G-1X RTT. The main conclusion is that end-to-end application-level performance is affected by various elements and layers of the network and thus it must be considered in all phases of the development process.

Keywords 3G Technology evolution · QoS · CDMA 2000 · Performance modeling · Simulator

J.-H. Kim (✉) · H.-J. Lee · S.-M. Oh
School of Electrical and Computer Engineering, AJOU University, 5 Wonchon-Dong Paldal-Gu, Suwon, Korea 442-749
e-mail: jkim@ajou.ac.kr

H.-J. Lee
e-mail: l33hyun@ajou.ac.kr

S.-M. Oh
e-mail: smallb01@ajou.ac.kr

S.-H. Cho
Digital Research Center, Samsung Advanced Institute of Technology, 14-1 Giheung-Eup Yongin-Si, Korea 449-712
e-mail: drcho@samsung.com

1 Introduction

The QoS issues in 3G CDMA wireless networks have been considered for many years. The standard committees such as 3GPP and 3GPP2 are working on QoS definition and provisioning in Universal Mobile Telecommunication Systems (UMTS) and Code Division Multiple Access (CDMA) 2000 networks. The 3GPP has defined the QoS architecture for four different QoS classes of traffics over UMTS networks. UMTS specifies four different QoS classes (or traffic classes) [1, 2]: Class 1 (Conversational), Class 2 (Streaming), Class 3 (Interactive), and Class 4 (Background). The main distinguishing factor among these classes is the sensitivity to delay. Conversational class is meant for those services that are very delay/jitter sensitive while Background class is insensitive to delay and jitter. Interactive and Background classes are mainly used to support traditional Internet applications like WWW, Email, Telnet, and FTP. Due to their looser requirements in delay as compared with Conversational and Streaming classes, both can achieve lower error rates by means of better channel coding and retransmission. The main difference between Interactive and Background classes is that the

former covers mainly interactive applications, such as web browsing and interactive gaming, while latter is meant for applications without the need of fast responses, such as file transferring or downloading of E-mails. 3GPP2 also uses these four classes in the CDMA2000 networks. The standard bodies focus the service class definition and QoS parameters for each class and each protocol layer only. However user perceived application-level performance can be obtained when the user application traffic transfers from client to server or mobile to mobile without any bottleneck or QoS violation point in the end-to-end reference connection. Thus QoS requirement should be mapped in every protocol layers in any network elements where the application packet pass by.

CDMA2000 3G-1X Radio Transmission Technology (RTT) was on the market from 2001. Many wireless service providers have been considering the 3G wireless technology migration path from circuit to packet technologies. Since the ATM transport can be used to support QoS in current technology but technology migration trends are looking for all IP transport in near future. Currently, 3GPP and 3GPP2 are working on the evolution of transportation technologies from ATM to IP to support integrated service and operation for internet access and voice traffic [3].

User performance studies for CDMA2000 were published in many papers [4–6]. In [4], data service performance were evaluated for 3G-1X RTT system but alternative architecture or voice service were not addressed. In [5], the TCP performance was presented in wireless interface but end-to-end performance was not included. Most papers addressed the wireless channel throughput or sector throughput and some others considered QoS strategies in CDMA2000 [6]. However, very few studies considered the whole network architecture. The user perceived application performance should be considered in an end-to-end reference architecture including a Radio Access Network (RAN), a Core Network (CN) and a data center, otherwise we can get only partial information on the application-level performance. To assess the user perceived application-level performance characteristics of different QoS service classes for alterative transport technologies and wireless technology evolution scenarios, we propose an end-to-end performance simulator for 2.5G or 3G-1X EV-DO networks. In this paper, we describe the end-to-end performance simulation model and methodology that we built for the CDMA 2000 network. We also address application-level performance issues in terms of wireless technologies evolution from 3G-1X RTT to 3G-1X EV-DO and transport technology evolution from ATM to IP. We modeled all the protocol layers from the physical through the application layer and modeled the detailed packet handling characteristics of each network element along the path. Foreground and background traffic loads are generated to represent specific application environment. The simulation model predicts application-level performance metrics such as response time, packet loss, jitter, and throughput. It is also used to assess architectural alternatives, identify performance bottlenecks, and validate performance requirements.

This paper is organized as follows. In Section 2, we introduce the proposed simulation methodology and the dynamic processor sharing strategy. In Section 3, we describe the detailed simulation model including the reference architecture, the protocol stack, and the service traffic models. In Section 4, we provide a detailed description of the simulation setup used for this study and the important simulation parameters used in Section 5. We then present some of the important simulation results in Section 5 and conclude this paper in Section 6.

2 Proposed simulation methodology and packet scheduling algorithm

2.1 Proposed simulation methodology

As mentioned above, to evaluate the user perceived application performance exactly, we should consider all network elements including a RAN, a CN and a data center. However, as the considered network elements and the service traffic loads increase, the simulation runtime is increased. The impact on the simulation run time, that is, precludes detailed application-level models. To solve this problem about the simulation runtime, we reviewed enormous number of different methods, such as statistical models for data traffic (long range dependant type) [7], [8] and traffic analysis and synthesis [9], [10]. Reyes-Lecuona proposed WWW traffic model using the hierarchical structure which consists of session, page, and packet level [7]. However, this algorithm is so complex that it can not use for network simulations where the number of events can be extremely large. Paxon proposed the empirically-derived analytic model for the data service traffic considered the various protocols using the traffic traced file [8]. But, it is also too complex to use the extremely large network simulation. Jain proposed the packet trains model described by a two-state Markov model [9]. It assumes that a group of packets travel together like a train. However, it is difficult that the packet trains model applies to the data service traffic which has a long-range dependence property since extremely many on/off sources are needed to generate a self-similar traffic. Lucas proposed a background traffic model designed for use in wide-area packet-switched network simulators [10]. The approach is to model the aggregate traffic generated by a large campus network and then partition the aggregated traffic into substreams, one for each campus-level destination in the backbone network. This model increases the computational efficiency to generate the self-similar traffic. However, this model is not considered of the user priority.

The proposed simulation methodology uses the packet traced file which records the packet size and the inter-arrival time. To calculate the impact according to the traffic load, we divide service traffic into separated service traffic model: the foreground service traffic and the background service traffic. The foreground service traffic is actually generated at the mobile terminal and can be used for the end-to-end application performance. However, the background service traffic is not generated at the mobile terminal and is generated at the each network elements using the traced file. It is used to calculate the service time for the foreground service traffic using the Lindley equation [11].

The first step is to collect a detailed packet trace for 1000 simultaneous application sessions using the simulator. This trace file is then scaled to match the desired mean rate for a given application. The trace file approach improved simulation run-time performance, but it was still too slow to run large scale network simulations. Thus, we used the trace file to simulate a virtual packet load by calculating the delay effect in the buffer instead of generating background traffic packet by packet. To calculate the packet delay effect, we used Lindley's recursion algorithm and extended it to account for the impact of multiple queues and queue scheduling disciplines. Lindley's recursion equation is given by

$$W_q^{(n+1)} = \begin{cases} W_q^{(n)} + S^{(n)} - T^{(n)} & \left(W_q^{(n)} + S^{(n)} - T^{(n)} > 0\right) \\ 0 & \left(W_q^{(n)} + S^{(n)} - T^{(n)} \le 0\right) \end{cases} \tag{1}$$

where, $W_q^{(n+1)}$ and $W_q^{(n)}$ mean waiting times of the $(n+1)$th packet and nth packet respectively. $S^{(n)}$ denotes the service time of the nth packet and $T^{(n)}$ means the inter-arrival time between the nth and $(n+1)$th packets. The packet delay calculation algorithm is as following.

- Definition
  - $F_0, F_1, \ldots, F_i, \ldots, F_p$: Background traced file with priority $0, 1, \ldots, F_i, \ldots, F_p$. 0 is highest priority and the F_i is the traced file for the current reference packet with priority i.
  - *t_last*: the time that the previous reference packet is arrived
  - *t_now*: the time that a current reference packet j arrived
  - *t_ia*: inter-arrival time between the previous packet and current testing packet k
  - *t_wait*: waiting time for the testing packet k which calculated by Lindley equation
  - *t_serv*: the service time for the previous packet $(k-1)$th for the testing packet k
  - *t_now'*: the departure time for the testing packet k. *t_now'* $=$ *t_now* $+$ *t_wait*
- Algorithm

Step 1. Calculate waiting time for the reference packet (priority i) for the F_i.

while (*t_last* + *t_ia* <= *t_now*) {
    t_wait = *t_wait* + *t_serv* − *t_ia* ;
    if (*t_wait* < 0)
        t_wait = 0;
    t_last = *t_last* + *t_ia*;
    }

Step 2. If there are any packet between *t_now* and *t_now*' in the any of higher priority background traffic traced file, then repeat 1 until no other higher priority traced packets are between *t_now* and *t_now.'*

Step 3. If there is any reference packet arrived between *t_now* and *t_now,'* defer the *t_now'* by the service time of reference packet(s) and re-calculate *t_now.'*

Step 4. Repeat step 1 until there is any other reference of background packet between *t_now* and *t_now.'*

2.2 Dynamic processor sharing (DPS) strategy

The proposed processor management strategy is essentially a DPS strategy, which utilizes a hybrid of priority and preemptive schemes for scheduling the processor in processing the bearer traffic with various QoS classes. We consider the four service classes which are defined in 3GPP because 3GPP2 use the service class defined 3GPP. The strategy uses the delay objectives of the different class in 3GPP standard [1, 2] for determining the appropriate share of processor real time for each corresponding class. 3GPP standard specifies the delay objectives for UMTS services as shown in Table 1. The Radio Access Bearer (RAB) delay tolerance is 80% of UMTS delay tolerance; I_u delay tolerance is 20% of RAB delay tolerance.

The processor time share assigned to each QoS class is based on the ratio of the delay tolerance of each class to the delay tolerance with respect to others, Let P_i be the share of processor time allocated to calss i. We have $\sum_{i=1}^{4} P_i = 1$, and $P_4 = 0$, given the four QoS classes defined in UMTS and Class 4 traffic is served with best effort. The radio bearer delay budget is then used to calculate the P_i. Let D_i be the delay budget for class i, we have

$$\begin{aligned} 1 &= P_1 + P_2 + P_3 \\ P_1 &= \frac{D_3}{D_1} P_3 \\ P_1 &= \frac{D_2}{D_1} P_2 \\ P_2 &= \frac{D_3}{D_2} P_3 \end{aligned} \tag{2}$$

Table 1 Delay requirements for UMTS QoS classes

	Conversation	Streaming	Interactive	Background
UMTS bearer RAB + CN bearer	100 msec	250 msec	400 msec	Best effort
RAB	80 msec	200 msec (80% of UMTS bearer)	320 msec (80% of UMTS bearer)	Best effort
CN bearer (SGSN to gateway)	20 msec	50 msec	80 msec	Best effort
I_u bearer	16 msec (20% of RAB)	40 msec (20% of RAB)	64 msec (20% of RAB)	Best effort
Radio bearer	64 msec (80% of RAB)	160 msec (80% of RAB)	256 msec (80% of RAB)	Best effort

Solving the above four equations with delay budget results in the following ratios, $P_1 = 0.61$, $P_2 = 0.24$ $P_3 = 0.15$, $P_4 = 0$, which implies that the share of processor time is allocated 61% for class 1, 20% for class 2 and 15% for class 3. Let T_i be the processor time assigned to class i, and C be the unit of processor time, we have

$$T_i = P_i \times C \tag{3}$$

With the above share assigned to each QoS class, a processor management strategy based on the priority as well as preemption schemes is proposed as following. Some similar strategy with Weighted Fair Queueing (WFQ) can be found in [12–16].

1. Processor is assigned to the processing Class 1 traffic whenever Class 1 queue is not empty.
2. When Class 1 traffic load becomes higher and the processor time spent in processing Class 1 traffic exceeded its share of T_1, stop accepting new call of Class 1.
3. If Class 1 queue is empty, then the processor is assigned to process the traffic in Class 2 and Class 3 queues by round robin manner with weighted share of T2 and T3.
4. Only when Class 1, 2, and 3 queues are all empty, the processor is assigned to serve Class 4 traffic.
5. In case of new traffic arrival at the queue of either Class 1 or 2 while Class 4 traffic is being processed, preemption of Class 4 processing is allowed.

Note that in Step 2, only new call of Class 1 is rejected while the traffic of existing call of Class 1 is protected and continues to have its highest priority in gaining the processor resources until the call is released. This is to guarantee the minimum delay and jitter in processing the Class 1 traffic due to its delay and jitter sensitivity as specified in 3GPP. The purpose of rejecting new calls of Class 1 when T_1 share is exceeded is to prevent the starvation of other lower level QoS classes such that they can also receive a fair share in processing that they deserve.

In Step 5, preemption is used to give a higher priority to the traffic of Class 1 and 2. This is also to minimize the delay or jitter for supporting the QoS of Class 1 and 2. On the other hand, preemption of Class 4 for a new arrival of Class 3 traffic is not needed. The minor gain in delay for Class 3 services (which are not so delay sensitive) may not be worthwhile as compared with the accompanied preemption overhead on the system. The preemption should not cause problem to Class 4 traffic because it is delay tolerance and is served in a best effort manner only. The preempted Class 4 traffic processing will be put at the head of the queue for Class 4, along with a tag indicating the remaining processing needed. As soon as the processor becomes available for Class 4, the preempted Class 4 traffic processing will be resumed and continued.

Throughout the whole process, the processor time spent in processing traffic of each QoS class needs to be monitored and accumulated. The actual share of each QoS class in processing time is derived from the record of accumulated time as needed. It is then used in Step 2 and 3 for comparison against the target share of T_1, T_2 and T_3 for determining the next traffic event to process accordingly. The concept of processor sharing among multiple queues of QoS classes is illustrated in Fig. 1.

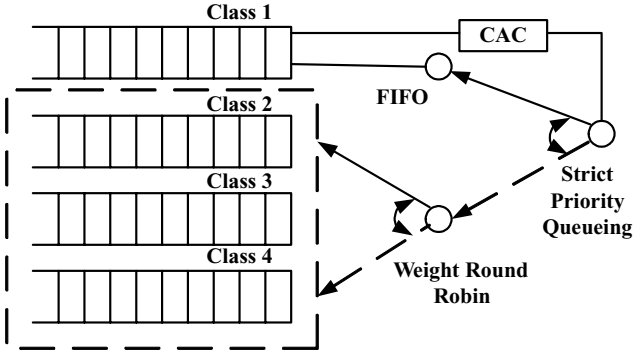


Fig. 1 Dynamic Processor Sharing (DPS) strategy for user traffic with QoS classes

3 Simulation model

3.1 Reference architecture and connection model

We study the performance modeling for the 2.5G and 3G-1X EV-DO networks. The 3G-1X system supports data rates from 9.6 kbps to 2.4 Mbps [17]. Figure 2 shows a reference network architecture model for the CDMA 2000. The reference network architecture can be considered into four different networks; RAN, CN, Internet and data center network. RAN may include Mobile Terminal (MT), Base Station Transmission System (BTS), Mobile Switching Center (MSC) and ATM or IP concentrators. CN includes ATM switches or IP routers and Packet Data Serving Nodes (PDSN). The data center network can be composed of three zones to protect servers from hacking or virus; a public, a DMZ (Demilitarized Zone), and a secure zone. Each zone can be protected by firewalls as shown in the Fig. 2.

We consider the data service and voice service reference connections on the reference network architecture. For a mobile web browsing service, a MT requests the data service to a web server in a data center. When the web server receives a web page request, the server returns back a response packet and transmits page information through the reference connection. The IP packets in a page are transferred from the server to the PDSN. The packets are then transmitted to the PCF (Packet Control Function) using GTP (GPRS Tunneling Protocol) tunnel through the CN which is ATM or IP networks. Most of current 3G-1X RTT networks use ATM and will be replaced by IP technology in the near future. PCF may be co-located inside MSC. The web page packets then are transmitted to the Serving Data Unit (SDU) in MSC.

For the voice traffic, a MT generates the appropriate codec packets (e.g. Code Exited Linear Prediction (CELP) or Enhanced Variable Rate Codec (EVRC)) and transmits them to InterWorking Function (IWF) in MSC. The IWF changes the speech coding in the cellular phone calls to the regular 64 kbps PCM, or 32 kbps ADPCM and sends it to CN.

3.2 Protocol architecture and model

In this paper, we consider two transport technologies such as ATM and IP in the RAN and CN. For ATM transport scenarios, a BTS chops a reverse link traffic packet into ATM cells and transmits them to MSC or RNC (for ALL IP scenario). Voice and data uses ATM Adaption Layer 2 (AAL2) and AAL5 in RAN respectively. For IP transport scenarios, BTS transmits an IP packet on the top of T1 and IP router converts it to Ethernet frame and sends it to MSC. The detailed protocol stack for ATM and IP protocol architectures are shown in Figs. 3 and 4. To assess the application-level performance, we implemented all the protocol stacks shown in Figs. 3 and 4 except the wireless channel model. To simulate the wireless channel error, we use the following link level simulation results. The channel model used in this paper is based on the models specified in 3G 1X-RTT. For link level simulation,

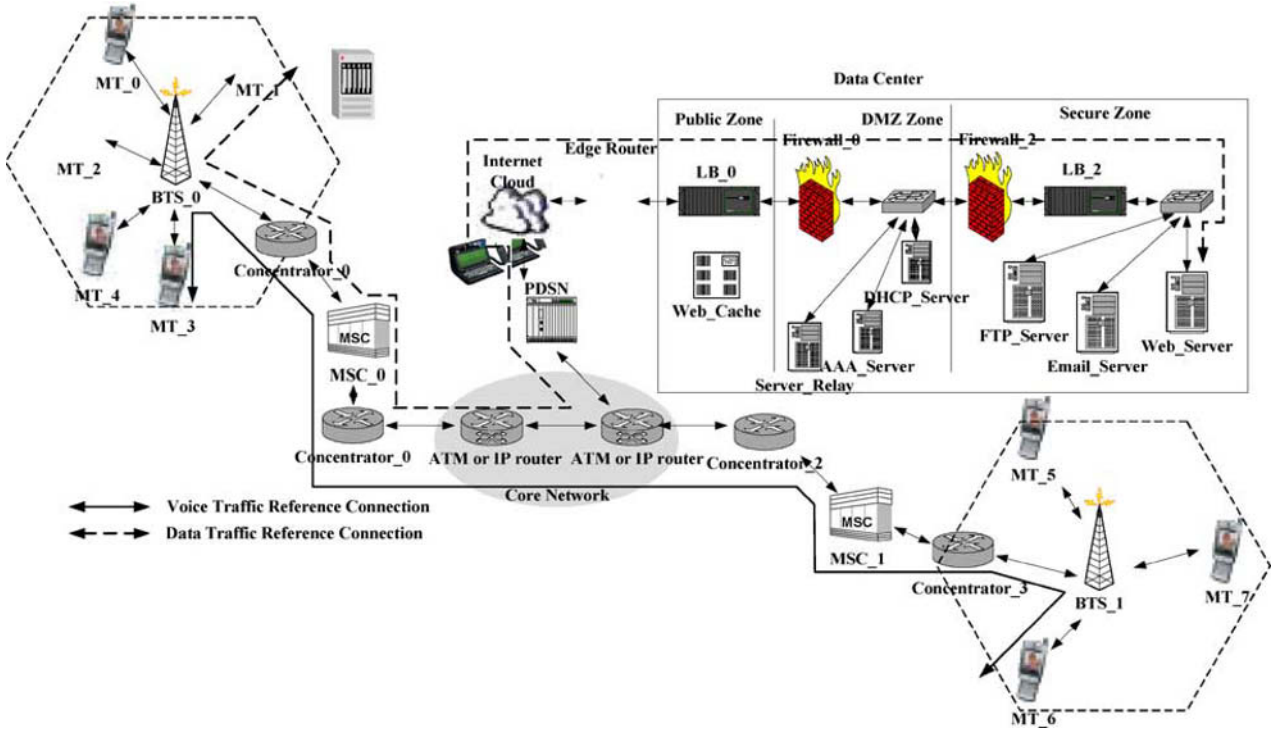


Fig. 2 Reference network model for the CDMA 2000

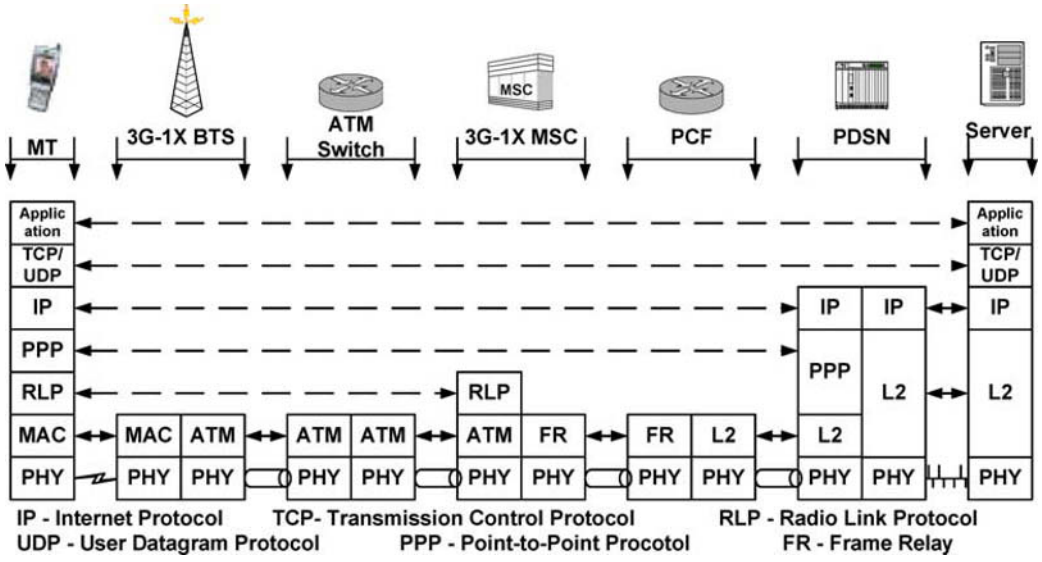
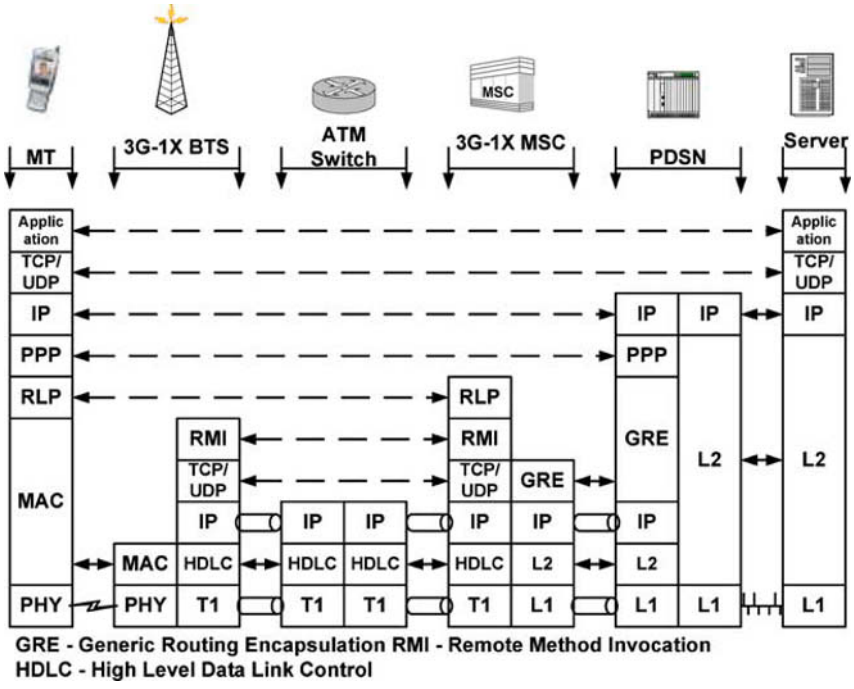


Fig. 3 Protocol stack model for the CDMA2000 using ATM transport architecture

Fig. 4 Protocol stack model for the CDMA 2000 using AII IP architecture



we use a traced file which contains frame error data when the target Frame Error Rate (FER) is fixed at 1%, 4% or 10%. These error data are the time co-related for each frame upon channel model.

The main protocols that affect application-level performance are MAC and Radio Link Protocol (RLP) in MT and MSC or Radio Network Controller (RNC). The CDMA2000 MAC sub-layer provides two important functions [18];

– Best Effort Delivery: reasonably reliable transmission over the radio link with a RLP that provides a "best effort" level of reliability
– Multiplexing and QoS Control: enforcement of negotiated QoS levels by mediating conflicting requests from competing services and by the appropriate prioritization of the access requests.

For the multiplexing and QoS control, MAC sub-layer is primarily responsible for determining the priorities among applications being run by several users and then determining the radio resources to them. Multiplex sub-layer is responsible for packing layer 2 SDUs onto one physical layer frame.

CDMA2000 system uses the RLP to recover the air interface frame error. The RLP uses the Negative

Acknowledgement (NAK) based Automatic Repeat Request (ARQ) mechanism. The RLP stores an a IP packet in a data buffer depending on the user session and fragments it to RLP frames. The RLP copies the RLP frame in the retransmit data buffer, and retransmits the particular RLP frame when the RLP receiver requests to retransmit it. When a RLP receiver detects the missing or corrupted RLP frame from a RLP sender, the RLP receiver sends the predetermined number of NAKs which specify the missing frame number to the RLP sender. (2,3) RLP means that RLP has two rounds to recover frame loss in the air. In the first round, whenever the RLP receiver detects a missing frame, it sends 2 NAKs to the RLP sender and activates the retransmission timer. If the missing frame is not received at the RLP receiver after the retransmission timer is expired, the RLP receiver then sends 3 NAKs in the second round to the RLP sender. In case the RLP receiver can not receive the missing frame after the second round, the RLP sends the imperfect IP packet to the upper layer. More detailed RLP protocol behavior can be found in IS-707 [19]. We fully implemented RLP protocol mechanism in the simulation model.

Another important factor affecting the application-level performance is packet schedulers in MSC or RNC. The packet scheduler in MSC (or RNC) decides data rates of a MT considering multiple factors such as Walsh code, remaining power, requested power requirement, buffer size, etc. 3G-1X EV-DO packet scheduling algorithms are different from those of 3G-1X RTT. The 3G-1X EV-DO standard does not specify any scheduling algorithm. In [20], it was suggested that the proportional fair scheduling is an appropriate compromise between maximizing system capacity and achieving fairness among users. This algorithm prioritizes user transmissions according to the value of the ratio *DRC/R*, where *DRC* (Data Rate Control) is the current value of the rate requested by the mobile on the DRC channel, and *R* is the IIR-filtered user throughput achieved by the mobile in the previous filter window.

3.3 Service traffic models

We use a hierarchical structure model to generate application level traffic. We use a voice traffic model and a web browsing traffic model for the applications in the paper. The voice traffic is generated by two hierarchical structures; call and packet level. The call level model composes of a sequence of ON and OFF periods as shown in Fig. 5. Each duration of ON and OFF period is exponentially distributed with mean 3 sec (activity factor is 0.5). During the ON period MT generates IS-733 13 kbps CELP and 8 kbps EVRC voice traffic packets [21, 22]. The CELP Codec processes 20 ms speech frame using four traffic packet types; Rate 1 (266 bits/packet), Rate1/2 (124 bits/packet), Rate 1/4 (54 bits/packet), and Rate 1/8 (20 bits/packet). The EVRC uses the Relaxed Code Excited Linear Prediction (RCELP) algorithm which is an analysis-by-synthesis algorithm that belongs to the class of speech coding algorithms known as CELP [22]. Unlike most existing CELP-type codecs, RCELP extracts pitch information once for every 20-ms speech frame, which results in larger algebraic codebook for prediction residuals and robustness under channel impairment conditions. The EVRC processes 20 ms speech frames using three of the four primary traffic packet types permitted by IS-95 Multiplex ; Rate 1 (171 bits/packet), Rate 1/2 (80 bits/packet) and Rate 1/8 (16 bits/packet). The rate is determined by the contents of the current speech frame and the history of the characteristics of the previous frames, as well as the command from the high level of the traffic controller. More details for the CELP and the EVRC can be found in [21] and [22]. We use the 3GPP2 standard traffic model [23] for a web browsing traffic. An example of the design model for a web browsing service is illustrated in Fig. 6. Here, we model the arrival of sessions. We also characterize the arrival of page requests within a session, and the number of objects and their sizes for each page. Detailed statistics can be found in Table 2 [23]. Other applications of interest can be modeled similarly.

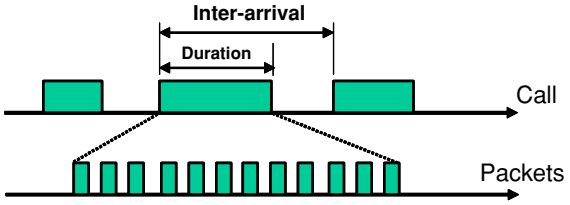


Fig. 5 Voice traffic model

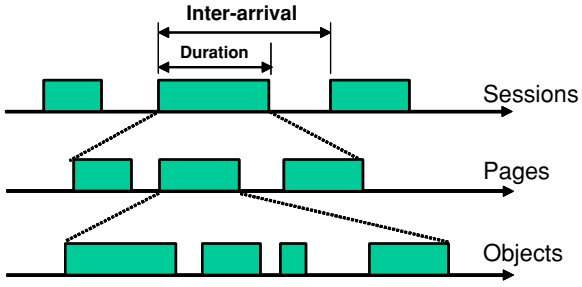


Fig. 6 Web browsing traffic model

4 Simulation scenarios and parameters

4.1 Simulation scenarios

In this study, we consider voice and data service scenarios. For a voice service scenario, we consider a mobile

Table 2 Simulation parameters

Category	Parameters	Reference
Voice traffic	8 kbps EVRC	[22]
Web browsing traffic	HTTP v.1.0 and 1.1	[23]
	Main object size: lognormal (10.8, 250) kbytes	
	Embedded object size: lognormal (7.8, 126) kbytes	
	Number of objects per page: Pareto shape :1.1 , location: 55	
TCP parameters	Windows 2000 based parameters	[23]
	MSS - 576 bytes	
	Window size - 16 kbytes	
	VJ header compression: No	
Radio link data rates (Kbps)	9.6, 153.6, 2000, 2400	[17]
Number of RLP round	5	[19]
RLP schemes	(2, 3) RLP scheme	[4, 19]
Voice coding scheme	IS-733 CELP (13 kbps), EVRC (8 kbps), G.711, G.726	
IP router processing time	100 micro sec	[26]
Frame error rates (FER)	0.01, 0.04, 0.1	
Processing time (msec)	MT - forward: 36.55, reverse: 63.05	[26]
	BTS - forward: 15, reverse: 9	
	MSC/RNC – forward: 7, reverse: 7	
	ATM/IP router: 0.1, Internet: 1.0	

to mobile voice call connection since it gives the worst performance. For the CN configuration, we use the 2.5G tandem backbone model, 3G ATM backbone and 3G IP backbone models. The detailed reference connection and reference network architecture model are shown in Fig. 7. To evaluate the application-level performance with different network configuration, we consider the following five scenarios:

– Scenario 1 (2.5G Tandem switch, 13 kbps voice codec): A voice packet is initiated from MT and transferred to BTS, the BTS then sends the voice packet on the ATM/AAL2 to MSC. In the MSC, IWF converts a voice packet to 64 kbps PCM packet format and sends it to a tandem switch. After the tandem switch, the voice packet is transmitted through PSTN and the other side of the network is symmetric.

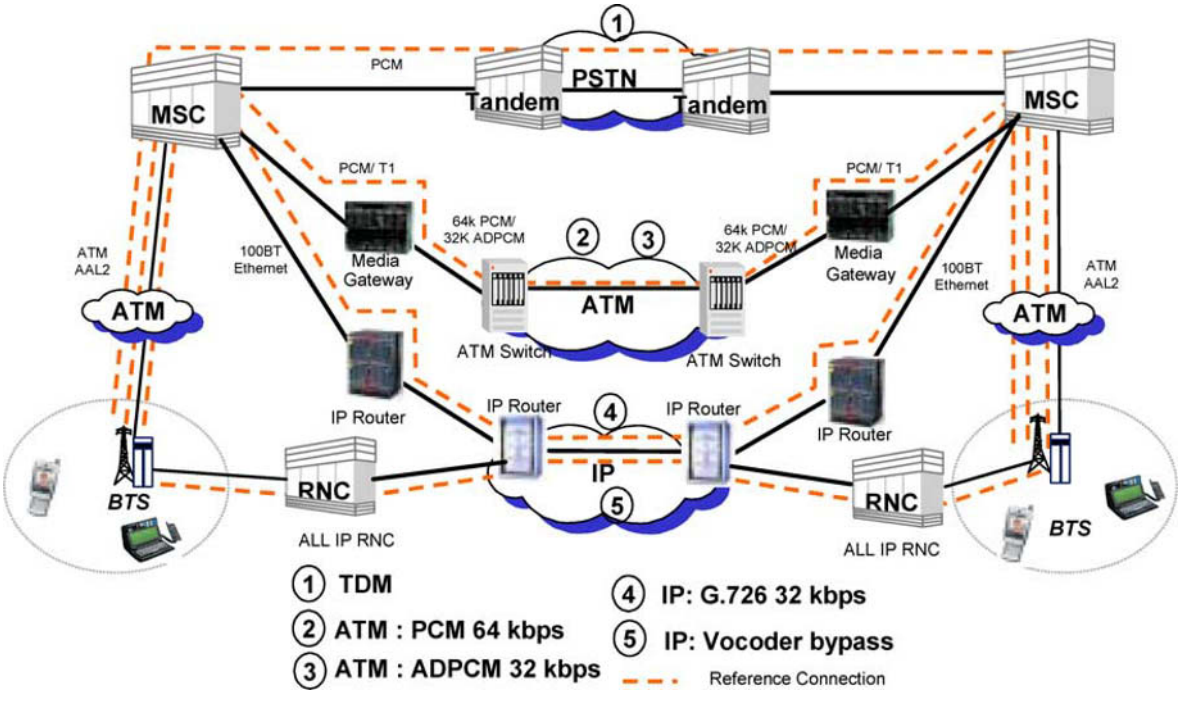


Fig. 7 Voice traffic reference connections and architectures

– Scenario 2 (3G ATM, G.711, 13 kbps voice codec): From MT to MSC is the same as Scenario1 but IWF in MSC converts a voice packet to a PCM packet format and sends it to a media gateway. The media gateway transfers PCM packet to ATM core network using ATM/AAL1.
– Scenario 3 (3G ATM, G.726, 13 kbps voice codec): From MT to media gateway is the same as Scenario 2. The media gateway converts a 64 kbps PCM packet to a G.726 32 kbps ADPCM packet and sends it to ATM core network using AAL2 multiplexing.
– Scenario 4 (3G IP, VoIP, 8 kbps EVRC): MSC converts an EVRC packet to a G.726 32 kbps packet and sends it to IP based media gateway using 100BT Ethernet. Then the IP media gateway sends it to IP CN.
– Scenario 5 (3G ALL IP, VoIP, 8 kbps EVRC, vocoder bypass): This is ALL IP scenario. An EVRC packet from MT is transmitted to BTS and IP based BTS sends it to IP RNC using 100BT Ethernet. The RNC then transfers the EVRC voice payload over an IP packet to IP CN. Header compression is also considered in the ALL IP scenario [15, 24]. We used the EVRC voice packet capsulated over PPP (4byte compressed header) and UDP/IP including compressed 3 bytes header from MT to IP router.

We consider 3 types of RAN transport technologies such as ATM, HDLC over T1 and 100BT Ethernet.

4.2 Simulation parameters

Two firewalls, two load balancers and 3 routers are modeled in a data center. For CN elements, media gateway, ATM switch and IP router models are implemented. We also implemented MSC (for 2.5G and 3G), RNC (for 3G-1X EV-DO), BTS, and MT models for RAN elements based on the 3 GPP specification release 4. The packet processing time for each network element follows the 3GPP standard specification [25]. We fully implemented each protocol in the network elements shown in Figs. 3 and 4.

RLP protocols in MT and MSC are implemented in accordance with the RLP3 specification of IS-2000. The air channel and physical layer is modeled based on the average channel quality and mobility. The user mobility model assumed that mobile users are uniformly distributed in a cell. Mobile users are assumed to move at a pedestrian speed of 3 km/h with worst case fading of single path Rayleigh. Based on the location of mobile terminal, we use the link level simulation result to estimate the power requirement for the user requested data rate. As described in Clause 3.2, we have implemented a 3G-1X RTT packet scheduler and a proportional fair scheduling algorithm for 3G-1X EV-DO scenario based on [20]. Some of simulation parameters are summarized in Table 2.

5 Simulation results and discussions

In this section, we assess the performance of the CDMA 2000 systems in terms of end-to-end delay of voice and data traffic service for 2.5G, 3G-1X RTT and 3G-1X EV-DO. We evaluate the performance of each service for different RAN architectures, CN architectures, and FERs. We show the simulation run time performance first and then show the simulation results for voice and data traffic service.

5.1 Simulation runtime performance results

The run-time performance of the simulation can be defined in terms of number of events and processing time per event. The simulation runtime performance is always an important issue, but is especially so for network simulations where the number of events can be extremely large. As mentioned previously, we have separated the traffic into the foreground and the background traffic and developed specialized techniques for handling each to improve the simulation efficiency.

To quantify the wireless simulator performance, we modeled FTP applications with varying numbers of users: The FTP application was a 1 Mbyte file download over the 64 kbps data rate and Table 3 shows some of the simulation run times. The simulation takes 120 sec for a single user. The simulation time increased linearly when the number of concurrent FTP sessions was added. It clearly shows that the simulation performance is not feasible when the number of concurrent application sessions is large. However, the last two rows of Table 3 show that the simulation performance is improved when additional ftp sessions are modeled as background traffic. For this scenario, 124 Mbps and 147 Mbps traffic on the average, which is about 80% and 95% of STM-1, was generated in all of the nodes along the reference connection (excluding application server) and one foreground FTP session was created.

There are two types of traffic modeled in the simulator: foreground and background traffic. The foreground traffic represents the specific services that traverse a given reference connection. The background traffic be representative of the applications that are being run over the network so that its

Table 3 Simulation run time with and without background traffic model

Number of foreground users	Number of background users	Download file size	Simulation time (sec)
1	0	1 Mbytes	120 sec
2	0	1 Mbytes	237 sec
3	0	1 Mbytes	355 sec
4	0	1 Mbytes	478 sec
5	0	1 Mbytes	596 sec
1	**1400**	**1 Mbytes**	**190 sec**
1	**1670**	**1 Mbytes**	**205 sec**

impact on the foreground traffic is accurately accounted for. It is the goal of the simulator to model the performance of the foreground traffic in detail. An important contributor to the performance of these services will be the characterization and load contributed by the background traffic at each node. In the simulation, background traffic is modeled at each network element by service type as if it arrived from a source and goes to a destination that is independent of the given reference connection.

5.2 Voice service performance simulation results

In this paper, we define the end-to-end voice packet delay as the sum of downlink delay, backbone delay, and uplink delay. To compare the end-to-end voice packet delay for the technology evolution from 2.5G to 3G-1X EV-DO, we perform the simulation for the five different scenarios. The first scenario is for 2.5G voice service, the second and the third scenarios are for current 3G-1X RTT systems, and the fourth and the last scenarios are for 3G-1X EV-DO systems. The results are presented in Fig. 8. The bottom part in the bar graph means uplink(reverse) delay which includes delay incurred in MT, BTS, MSC (or RNC) for voice packet. The middle block means the CN packet delay which includes ATM or IP router processing delay, and the top of the bar is the down link (forward) packet delay which means the delay from MSC (or RNC) to MT. The background traffic load for each network element is 40% in the simulation. In the second and the third scenarios (ATM CN), voice packet delay is a little bit larger than that of the first case (tandem switch). Because in both scenarios ATM processing delay is included in the CN, and the AAL2 multiplexing delay is (processing delay and Timer_CU : 2 msec) also included in CN for the third scenario. In the fourth scenario (IP CN, G.726 ADPCM), the CN packet delay is increased compared with the AAL1 scenario (Scenario 2) since the IP packet overhead for voice traffic is larger than packet format overhead of ATM. The last scenario is for the vocoder bypass which means that an EVRC voice packet is transferred over the IP packet without any trans-coding to other coding scheme. It reduces RAN and CN coding processing delay and results into 32% delay reduction compared with the third scenario. If we map one-way end-to-end delay to the R value in E model in [27], the voice quality for the first and the second scenarios provide high quality voice. Scenario 3 and Scenario 4 provide medium quality voice, while the vocoder bypass scenario meets the high quality voice. From the simulation, we observed that, in the voice user's performance view, IP transport in CN does not provide much performance improvement over ATM. In the vocoder bypass scenario, the packet delay decreases significantly and as a result voice quality improves noticeably. If we consider the error generated in coding and transcoding from EVRC to other coding schemes, the vocoder bypass scenario is the best solution to support VoIP in CDMA2000 systems.

To measure the impact of the data traffic load increase to the voice quality, we increase the background traffic load in every network element from 0% to 80% when the forward and reverse data rates are fixed at 153 kbps and 64 kbps respectively. The mean voice end-to-end delay and the mean jitter according to the service traffic load are presented in Figs. 9(a) and (b), respectively. All of the simulation is not used IP QoS mechanism. In simulation results, we know that the mean end-to-end delay and the jitter are increased as the

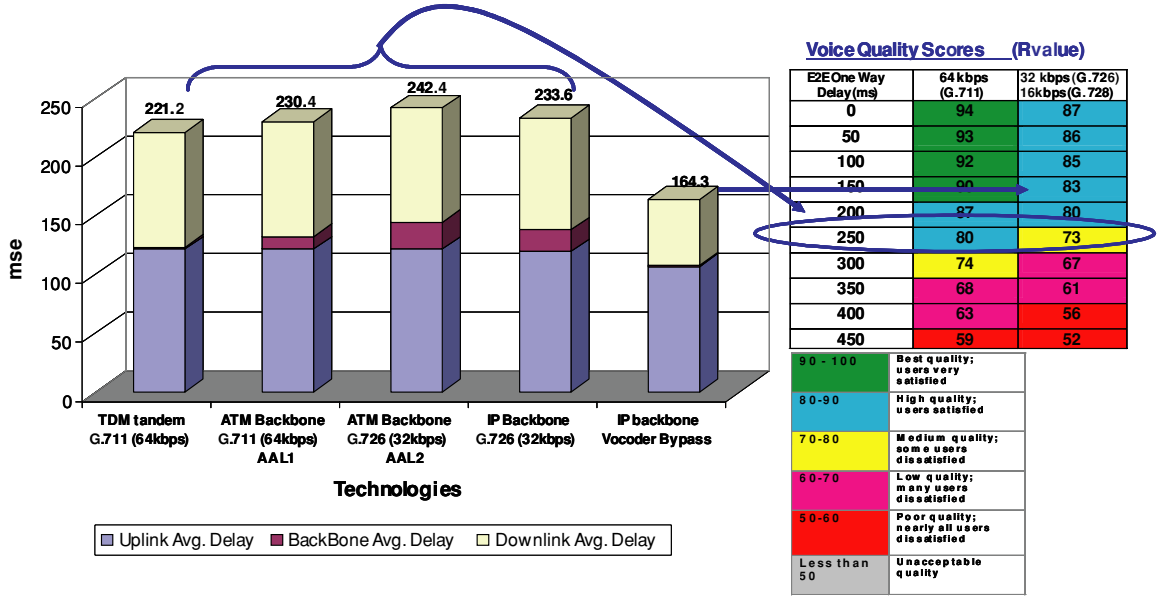


Fig. 8 The end-to-end voice packet delays for technology evolution

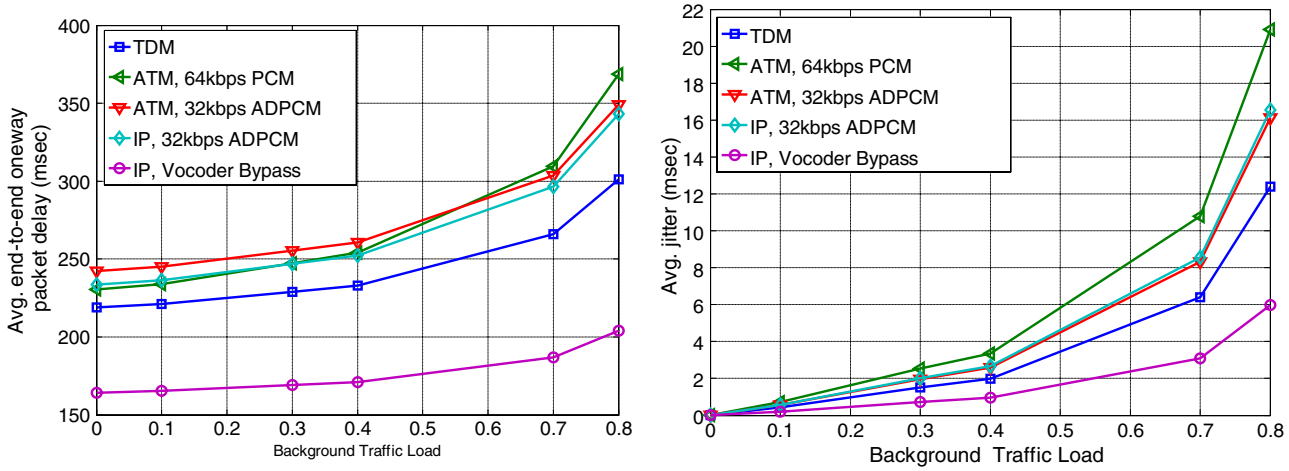


Fig. 9 The average end-to-end packet delay and the average jitter vs. the background traffic load

background traffic load increases. In Fig. 9, the mean delay in Scenario 2 is crossing the mean delay in Scenario 3 and Scenario 4 when the service traffic loads are 30% and 54%. The transcoding delay is the major component for the delay at the low traffic load. However, the packet service time is more important element for the delay as traffic load increases. When the data traffic load is 70% except Scenario 5, the end-to-end voice packet delay exceeds the voice delay requirement which is around 250 msec [26]. Thus, to meet the end-to-end voice quality requirement, RAN and CN must have IP QoS mechanisms.

We evaluate the performance of the proposed packet scheduling scheme, DPS according to the service traffic load in Fig. 10. We assumed that the weighting factors for service traffics are 0.61, 0.24, 0.15, and 0, respectively. DPS algorithm guarantee voice services to 61% grant to satisfy the maximum delay for each service classes. Therefore, DPS suppress the rate of increase of the delay as the service traffic load increases.

5.3 Data service performance simulation results

5.3.1 HTTP v.1.0 vs. HTTP v.1.1

HTTP (Hyper Text Transport Protocol) is the main protocol for web browsing service. HTTP version 1.1 is a modified version of HTTP v.1.0. The main differences are the persistent connection and the pipelining. HTTP v.1.0 uses a TCP connection for every single object in a web page, but a single TCP connection can be used to multiple object transmission in the persistent connection, which means that the page response time can be reduced by eliminating multiple TCP three-way handshaking and slow starts. HTTP v.1.0 sends a request packet for every single object. However, HTTP v.1.1

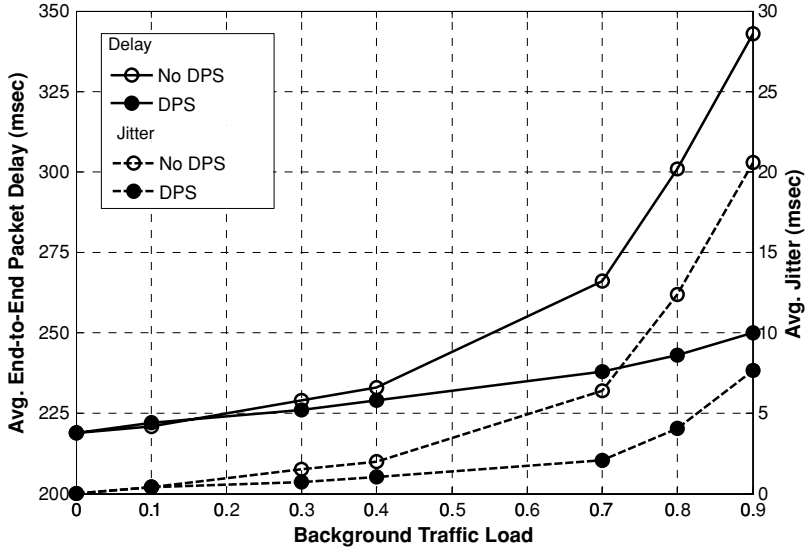


Fig. 10 Avg. end-to-end packet delay and avg. jitter with and without DPS

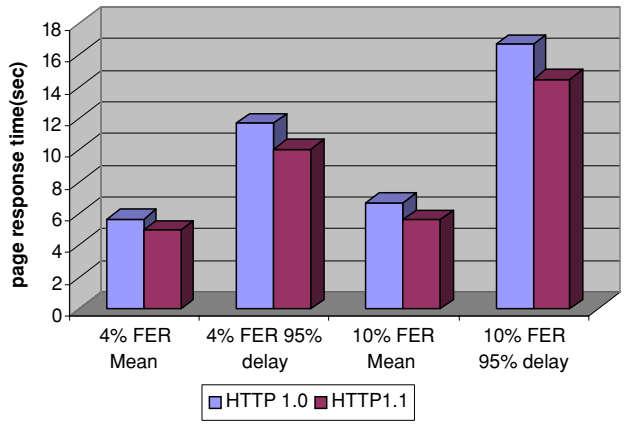


Fig. 11 Page response time for HTTP c.1.0 vs. HTTP v.1.1

uses pipelining which uses a single request packet for multiple objects so that it reduces the request packet transfer delay.

We used the third scenario to simulate HTTP performance. Figure 11 shows the mean and the 95 percentile page response time for different FERs. As expected, HTTP v.1.1 shows shorter response time for all FERs. HTTP v.1.1 shows 12.3% and 16.7% reduced page response delay for 4% and 10% FERs respectively. As the FER increases, the page response time also increases. Because, in both HTTP versions, the number of TCP retransmissions for object request packets and objects are increased for the higher FER even though RLP recover some of air frame errors.

5.3.2 ATM vs. IP in radio access network

ATM transport technology in current 3G network will eventually migrate to IP technologies. To compare ATM and IP transport technologies in RAN, we measure the web page response time using the same scenario as in the HTTP performance evaluation. Figure 12 presents the web page response time for three different RAN transport technologies: ATM, HDLC over T1, and 100BT Ethernet. When FER is low, there is no significant difference of delay performance among the three RAN transport technologies. However, when FER is high, the difference of delay performance is significant. At 10% FER, we observe 15.6 and 21.7% page response time reduction when RAN transport technology migrates from ATM to HDLC and ATM to 100 BT Ethernet, respectively. The 21.7% performance improvement is due to the higher transmission speed and lower packet overhead in IP layer. In this case, IP transport technology is better solution for higher FER environment since the IP packet overhead is smaller than that of ATM for web browsing data traffic, while it shows the opposite effect to small size voice packets as shown in Fig. 8.

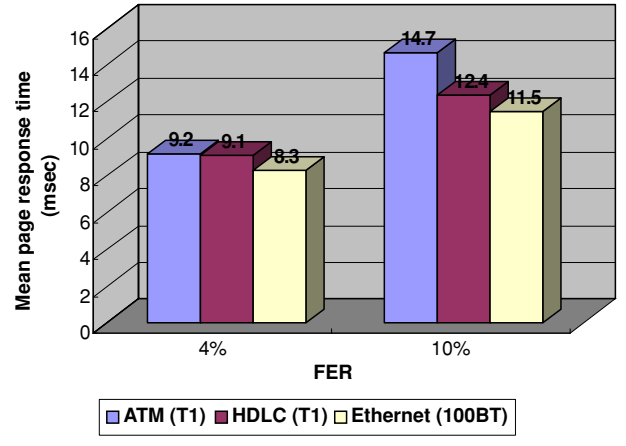


Fig. 12 Web page response time for different RAN transport technologies

5.3.3 ATM vs. IP in core network

Until IP technology becomes mature enough to support QoS, ATM transport will be used in CN. The major advantages of IP technologies are simplicity and easy management with unified management systems in supporting different services. Figure 13 presents the web page response time when either ATM or IP is adopted as a CN technology. Delay performance improvement in IP CN technology is about 5.2 to 7.7% compared with to ATM. In the delay performance aspect, IP does not provide much improvement. This simulation implies that once ATM infrastructure is in place, the evolution to IP CN may not be urgent, since the performance improvement is marginal compared with the huge investment it may involve.

5.3.4 3G-1X RTT vs. 3G-1X EV-DO

3G-1X EV-DO service started from 2001 in Korea. 3G-1X EV-DO enhances the data rate to 2.4 Mbps. To compare the

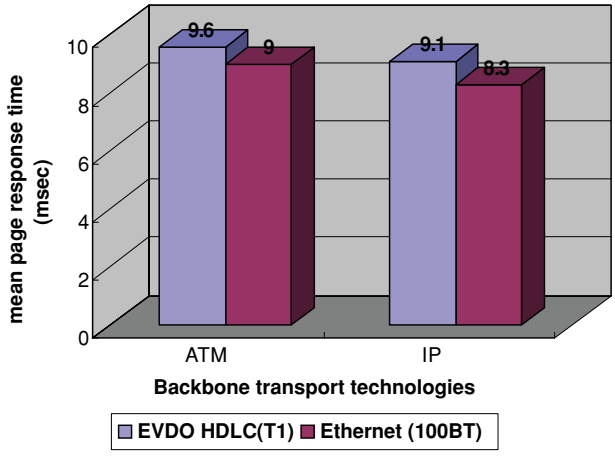


Fig. 13 Web page response time for ATM and IP CN transport technologies

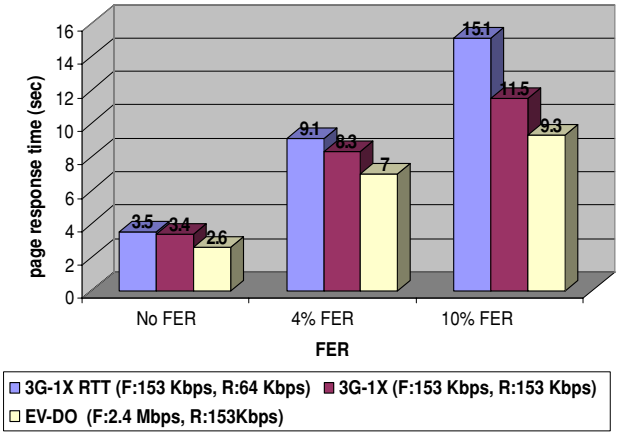


Fig. 14 Web page response time for 3G technology

data service performance for 3G wireless technology, we measure the web browsing response time for different data rates in 3G-1X RTT and 1X EV-DO networks. Figure 14 presents the web page response time according to the different wireless technologies. We assume that 100 BT Ethernet RAN and IP CN transport technology in this scenario. There is no significant performance difference when the channel is error free. However, at 10% FER, EV-DO provides 38% reduced response time compared to 3G-1X RTT. The higher data rate in RAN of EV-DO results in faster frame retransmission compared with 3G-1X RTT. As FER increases, the performance difference between the two technologies becomes more significant.

6 Conclusions

In this paper, we described the end-to-end performance simulation model and methodology that we had built for the CDMA 2000 network. We also addressed application-level performance issues in terms of wireless technologies evolution from 2.5G to 3G-1X EV-DO and transport technology evolution from ATM to IP. The simulator modeled all protocol layers from the physical to the application layers and modeled details of the packet handling characteristics of each network element along the path. The foreground and background traffic loads were generated to represent specific applications running over the network. In the simulation we considered the end-to-end reference architecture and connection including RAN, CN, and data center.

For the voice service, we found that if we consider the error generated in coding and transcoding from EVRC to other coding schemes, the vocoder bypass scenario is the best solution to support VoIP in CDMA2000 systems. We also found that end-to-end QoS mechanism should be provided in every network element where the packet passes by.

For data packet performance, we found that HTTP v.1.1 shows better performance than that of HTTP v.1.0 due to the pipelining and TCP persistent connection. We also found that IP transport technology is better solution for higher FER environment since the IP packet overhead is smaller than that of ATM for web browsing data traffic, while it shows opposite effect to small size voice packet in RAN architecture. We captured the page response time for different 3G technologies with different data rates. The 3G-1X EV-DO system with 2.4 Mbps forward link and 153 Kbps reverse link data rates showed 38% delay reduction compared to 3G-1X RTT system. Though this performance improvement does not show all the performance aspects of the 3G-1X EV-DO system, we can imagine the performance improvement that the 3G-1X EV-DO network may give.

Through the simulation analysis, we found that the performance advantages of IP Core Network over ATM may not big enough to justify the evolution from ATM to IP once ATM network is already in place. However, when other factors such as reliability, expenditure costs and operation/management costs are considered, evolution to All IP network may be justified.

Acknowledgments The authors would like to thank the anonymous reviewers for their comments and suggestions which helped to improve the presentation of the paper. This work was supported by the Ministry of Education (MOE) of Korea by its second stage of Brain Korea 21(BK21) Program.

References

1. 3GPP TS 22.105 V4.3.0, *Services and Service Capability Release* Vol. 4 (March 2002).
2. 3GPP TS 23.107 V.4.5.0, *Quality of Service Concept and Architecture Release* Vol. 4 (September 2002).
3. 3GPP TS 23.207 V.5.8.0, *End-to-End Quality of Service Concept and Architecture* (June 2003).
4. Z. Dziong, F. Khan, K. Medepalli and S. Nanda, Wireless internet access using IS-2000 third generation system: a performance and capacity study, Wireless Networks Vol. 8 (August 2002) pp. 325–336.
5. E. Chaponniere, S. Kandukuri and W. Hamdy, Effect of physical layer bandwidth variation on TCP performance in CDMA2000, in: *Proc. of IEEE Vehicular Technology Conference 2003 Spring* Vol. 1 (April 2003) pp. 336–342.
6. F. Li, M. Nguyen and W. Seah, QoS support in IP/MPLE-based radio access networks, in: *Proc. of IEEE Vehicular Technology Conference 2003 Spring* Vol. 4 (April 2003) pp. 2251–2255.
7. A. Reyes-Lecuona, et al., A page-oriented WWW traffic model for wireless system simulations, in: *Proceedings of ITC-16*, Edinburgh, United Kingdom (June 1999) pp. 1271–1280.
8. V. Paxson, Empirically-derived analytic models of wide-area TCP connections, IEEE/ACM Trans. on Networking Vol. 2, Issue 4 (August 1994) pp. 316–336.
9. R. Jain and S. Routhier, Packet train-measurements and new model for computer network traffic, IEEE JSAC Vol. 4, Issue 6 (September 1986) pp. 986–995.

10. M.T. Lucas, B. Dempsey, D. Wrege and A. Weaver, An efficient self-similar traffic model for wide-area network simulation, in: *Proc. of IEEE GLOBECOM '97*, Phoenix, Arizona, USA Vol. 3 (November 1997) pp. 1572–1576.
11. L. Kleinrock, *Queueing Systems* (Wiley, New York, 1975).
12. M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar and R. Vijayakumar, Providing quality of service over a shared wireless link, IEEE Commun. Mag. (February 2001) 150–154.
13. H. Zhang, Service disciplines for guaranteed performance service in packet-switching networks, in: *Proceedings of the IEEE* Vol. 83, Issue 10 (October 1995) pp. 1374–1396.
14. A. Parekh and R. Gallager, A generalized processor sharing approach to flow control in integrated services networks: the single node case, IEEE/ACM Trans. on Networking Vol. 1 (June 1993) pp. 139–150.
15. A. Demers, S. Keshav and S. Shenkar, Analysis and simulation of a fair queueing algorithm, in: *Proc. of ACM SIGCOMM*, Austin, TX, USA (September 1989) pp. 1–12.
16. O. Sallent, J. Romero, R. Agusti and F. Casadevall, Provisioning multimedia wireless networks for better QoS: RRM strategies for 3G W-CDMA, IEEE Commun. Mag. Vol. 41, Issue 2 (February 2003) pp. 100–106.
17. 3GPP2 S.R 0023 V.2.0, *High-Speed Data Enhancements for cdma2000 1X Data Only* (November 2000).
18. 3GPP2 C.S 0003-C V.1.0, *MAC Standard for cdma2000 Spread Spectrum Systems release C* (February 2002).
19. TIA/EIA/IS-707-A-1.10, *Data Service Options for Spread Spectrum Systems: Radio Link Protocol Type 3* (December 1999)
20. A. Jalali, R. Padovani and R. Pankaj, Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system, in: *Proc. of IEEE Vehicular Technology Conference*, Tokyo, Japan Vol. 3 (May 2000) pp. 1854–1858.
21. 3GPP2 C.S0020-0, *High Rate (13 kbps) Speech SO* (December 1999).
22. 3GPP2 C.S0014-0 V.1.0, *Enhanced Variable Rate Codec* (December 1999).
23. 3GPP2 TSG-C.R1002, *1XEV-DV Evaluation Methodology (V13)* (2003).
24. V. Jacobson, IP headers for low-speed serial links, RFC 1144, Internet Engineering Task Force (Feb. 1990).
25. 3GPP2 S.R. 0069, V. 1.0, *Header Stripping and Generation* (March 2002).
26. 3GPP TR 25.853 v4.0.0, *Delay Budget within the Access Stratum* (March 2001).
27. M. Perkins, C. Dvorak, B. Lerich and J. Zebarth, Speech transmission performance planning in hybrid IP/SCN networks, IEEE Commun. Mag. (July 1999) pp. 126–131.

Jae-Hyun Kim He received the B.S., M.S., and Ph.D. degrees, all in computer science and engineering, from Hanyang University, Ansan, Korea, in 1991, 1993, and 1996 respectively. In 1996, he was with the Communication Research Laboratory, Tokyo, Japan, as a Visiting Scholar. From April 1997 to October 1998, he was a post-doctoral fellow at the department of electrical engineering, University of California, Los Angeles. From November 1998 to February 2003, he worked as a member of technical staff in Performance Modeling and QoS management department, Bell laboratories, Lucent Technologies, Holmdel, NJ. He has been with the department of electrical engineering, Ajou University, Suwon, Korea, as an assistant professor since 2003. His research interests include QoS issues and cross layer optimization for high-speed wireless communication. Dr. Kim was the recipient of the LGIC Thesis Prize and Samsung Human-Tech Thesis Prize in 1993 and 1997, respectively. He is a member of the Korean Institute of Communication Sciences (KICS), Korea Institute of Telematics and Electronis (KITE), Korea Information Science Society (KISS), and IEEE.

Hyun-Jin Lee received the B.S. degree in electrical engineering from Ajou University, Suwon, Korea, in 2004, and is working toward the M.S. degree and Ph. D. degree in electrical engineering at Ajou University. He has been awarded Samsung Human-Tech Thesis Prize in 2004. His research interests QoS, especially network optimization and wireless packet scheduling. He is a member of the KICS.

Sung-Min Oh received the B.S. and M. S. degrees in electrical engineering form Ajou University, Suwon, Korea, in 2004, and is working toward the Ph. D. degree in electrical engineering at Ajou University. His research interests QoS performance analysis and 4G network. He is a member of the KICS.

Sung-Hyun Cho received his B.S., M.S., and Ph.D. in computer science and engineering from Hanyang University, Korea, in 1995, 1997, and 2001, respectively. From 2001 to 2005, he has been with Samsung Advanced Institute of Technology, where he has been engaged in the design and standardization of MAC and upper layers of B3G, IEEE 802.16e, and WiBro systems. He is currently a MAC part leader in the telecommunication R&D center of Samsung Electronics. His research interests include 4G air interface design, radio resource management, cross layer design, and handoff in wireless systems.