Internet Engineering Task Force
INTERNET DRAFT
Signaling Transport Working Group
October 22, 1999
Expires March 22, 2000

Authors
Huai-An P. Lin
Kun-Min Yang
Taruni Seth
Christian Huitema
Telcordia Technologies

VoIP Signaling Performance Requirements and Expectations
<draft-ietf-sigtran-performance-req-01.txt>

Status of this document

This document is an Internet-Draft and is in full conformance with all provisions of Section 10 of RFC2026. Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html.

Abstract

This document serves as input into the IETF SIGTRAN requirements process. It includes call setup delay requirements, derived from relevant ISDN and SS7 standards published by ITU-T (International Telecommunications Union--Telecommunications Standardization Sector) and generic requirements published by Telcordia Technologies (formerly Bellcore). To gain user acceptance of Voice-over-IP (VoIP) services and to enable interoperability between Switched Circuit Networks (SCNs) and VoIP systems, it is imperative that the VoIP signaling performance be comparable to that of the current SCNs. The requirements given in this Internet Draft are intended to be the worst-case requirements, for at least in the United States SCN calls are typically set up at a faster speed than the derived requirements. At the end of the draft, several VoIP call connection scenarios based on the latest megaco protocol are analyzed and compared with similar cases in the PSTN. It indicates the PDD performance of VoIP systems is somewhat worst but not by much. An

improvement in some network element can bring VoIP systems to have comparable PDD performance as the PSTN.

1. Introduction

This document serves as input into the IETF SIGTRAN requirements process. It includes call setup delay requirements, derived from relevant ISDN and SS7 standards published by ITU-T (International Telecommunications Union--Telecommunications Standardization Sector) and generic requirements published by Telcordia Technologies (formerly Bellcore). To gain user acceptance of VoIP services and to enable interoperability between SCNs and VoIP systems, it is imperative that the VoIP signaling performance be comparable to that of the current SCNs. The requirements given in this Internet Draft are intended to be the worst-case requirements, since at least in the United States SCN calls are typically set up within one to two seconds [1]--far faster than the derived requirements.

The call setup delay, also known as the Post Dial Delay (PDD), in an ISDN-SS7 environment is the period that starts when an ISDN user dials the last digit of the called number and ends when the user receives the last bit of the Alerting message. Call setup delays are not explicitly given in the existing SCN performance requirements; rather, performances of SCNs are typically expressed in terms of cross-switch (or cross-office) transfer times. This Internet Draft uses ITU-T's SS7 Hypothetical Signaling Reference Connection (HSRC) [2], cross-STP (Signaling Transfer Point) time [3], Telcordia's switch response time generic requirements [4], and a simple ISDN-SS7 call flow to derive the call setup delay requirements. ITU-T's cross-switch time requirements [5] are listed as references but not used, since the ISDN timings are missing.

At the end of the draft, we evaluate the PDD of VoIP systems based on the proposed megaco protocol and compare its PDD performance with that of the current PSTN. It gives a better understanding of where the bottleneck is and hopefully suggest the area of improvement that can be done in VoIP systems to achieve comparable performance.

2. Hypothetical Signaling Reference Connection (HSRC)

HSRC is specified in ITU-T Recommendation Q.709. A HSRC is made up by a set of signaling points and STPs that are connected in series by signaling data links to produce a signaling connection. Recommendation Q.709 distinguishes the national components from the international components. A HSRC for international working consists of an international component and two national components. The size of each country is considered; however, the definitions of large and average countries was not precisely defined:

When the maximum distance between an international switching center and a subscriber who can be reached from it does not exceed 1000 km or, exceptionally, 1500 km, and when the country has less than n x 10E7

subscribers, the country is considered to be of average-size. A country

with a larger distance between an international switching center and a subscriber, or with more than n x 10E7 subscribers, is considered to be of large-size. (The value of n is for further study.)

Recommendation Q.709 uses a probabilistic approach to specify the number of signaling points and STPs on a signaling connection. The maximum number of signaling points and STPs allowed in a national component and an international component are listed in Tables 1 and 2, respectively.

Table 1: Maximum Number of Signaling Points and STPs in a National Component (Source: ITU-T Recommendation Q.709, Table 3)

Country size	Percent of connections	Number of STPs	Number of signaling points*
Large-size	50%	3	3
	95%	4	4
Average-size	50%	2	2
	95%	3	3

* The terms signaling points and switches are used interchangeably in this Internet Draft.

Table 2: Maximum Number of Signaling Points and STPs in International Component (Source: ITU-T Recommendation Q.709, Table 1)

Country size	Percent of connections	Number of STPs	Number of signaling points
Large-size to Large-size	50%	3	3
	95%	4	3
Large-size to Average-size	50%	4	4
	95%	5	4
Average-size to Average-size	50%	5	5
	95%	7	5

3. Switch Response Time (aka Cross-switch Transfer Time)

Most of SCN performance requirements are specified in terms of switch

response times, which are also referred to as cross-switch transport time or cross-switch delay. This section reviews the meanings of switch response times, several other related terms, and the generally accepted values of switch response times published by Telcordia Technologies. The corresponding ITU-T's cross-switch timing requirements are also listed as references.

This Internet Draft reviews the switch response time requirements intended to apply under normal loading. Normal loading is usually associated with the notion of the Average Busy Season Busy Hour (ABSBH) load. Simply put, it is expected that the switch response times that a particular switch experiences at this load will be virtually load-independent.

Switch response time is the period that starts when a stimulus occurs at the switch and ends when the switch completes its response to the stimulus. The occurrence of a stimulus often means the switch receives the last bit of a message from an incoming signaling link, and completion of a response means the switch transmits the last bit of the message on the outgoing signaling link. If the switch's response to a stimulus involves the switch sending a message on the outgoing signaling link, then switch processing time is the sum of the switch processing time and the link output delay:

Switch Response Time = Switch Processing Time + Link Output Delay

Switch processing time is the period that starts when a stimulus occurs at the switch and ends when the switch places the last bit of the message in the output signaling link controller buffer. The period between the switch placing the message in the output signaling link controller buffer and the switch transmitting the last bit of the message on the outgoing signaling link is defined as the link output delay. Link output delay can be further divided into the queuing delay and message emission time. There are separate delay requirements for switch processing time and link output delay; however, for simplicity only the combined delay requirements for switch response time, as given in Table 3, will be listed in this Internet Draft.

Table 3: Switch Response Time Assuming Typical Traffic Mix and Message Lengths (Source: Telcordia GR-1364-CORE, Table 5-1)

Type of Call Segment	Switch Response Time (ms)	
	Mean	95%
ISUP Message	205-218	<=337-349
Alerting	400	<=532
ISDN Access Message	220-227	<=352-359
TCAP Message	210-222	<=342-354
Announcement/Tone	300	<=432
Connection	300	<=432
End MF Address - Seize	150	<=282

Telcordia GR-1364 specifies switch response time using switch call segments as a convenient way to refer to the various phases of call processing that switches are involved in. (An alternative would be proposing switch processing requirements for every possible type of switch processing. Obviously, this would become burdensome and would necessitate adding to the requirements every time an additional type of switch processing was required.) Listed in Table 3 are:

1. **ISUP message call segments that involve the switch sending an ISUP message as a result of a stimulus.**
2. **Alerting call segments that involve the switch alerting the originating and/or terminating lines as a result of a stimulus.**
3. **ISDN access message call segments that involve the switch sending an ISDN access message (other than an ISDN access ALERT message) as a** result of stimulus. ISDN access message call segment processing occurs at originating or terminating switches where the originating or terminating line, respectively, is an ISDN line.
4. **TCAP message call segments that involve the switch sending a TCAP message as a result of a stimulus.**
5. **Announcement/tone call segments that involve the switch playing an announcement, placing a tone on, or removing a tone from the** originating or terminating line as a result of a stimulus. However, the announcement/tone call segments do not include dial-tone delay, of which the delay requirements can be found in Telcordia TR-TSY-000511[6].
6. **Connection call segments involve the switch connecting one or more users as a result of a stimulus.**

The ITU-T's cross-switch timing requirements are listed below as references. It is noted that the ITU-T's requirements are noticeably stringent that those of Telcordia under the normal loading. However, since the ITU-T's values are stated as provisional and they do not provide the timing requirements for ISDN, Telcordia's values will be used to derive the call setup delay requirements.

Table 4: ITU-T Cross-Switch Transfer Time
(Source: ITU-T Recommendation Q.725, Table 3)

Message type	Exchange call attempt loading	Cross-Switch Transfer Time (ms)* Mean	95%
Simple (e.g. answer)	Normal	110	220
	+15%	165	330
	+30%	275	550
Processing intensive (e.g. IAM)	Normal	180	360
	+15%	270	540
	+30%	450	900

* Provisional values.

4. Cross-STP Delay

Message delay through an STP is specified as the cross-STP delay. It is the interval that begins when the STP receives the last bit of a message from the incoming signaling link, and ends when the STP transmits the last bit of the message on the outgoing signaling link. As with the switch response time discussed in the previous section, the cross-STP can be divided into processor handling time and link output delay. This Internet Draft adopts the cross-STP delay requirements specified in ITU-T Q.706 Recommendation.

Table 5: Message transfer time at an STP
(Source: ITU-T Recommendation Q.706, Table 5)

STP signaling traffic load	Message transfer Time (ms) Mean	95%
Normal	20	40
+15%	40	80
+30%	100	200

5. Maximum End-to-End Signaling Delays

Using the HSRC, switch response times, and cross-STP delays, one can compute the maximum signaling transfer delays for ISUP messages under normal load. As with Telcordia GR-1364, it is assumed that the distribution of switch response time for each call segment is approximately a normal distribution. It is further assumed that switch response times of different switches are independent. Under these assumptions, the end-to-end (from originating switch to terminating

switch) delays for each national component and for international calls are listed in Tables 6 and 7, respectively. The 20 ms cross-STP delay is assumed in all cases. It should be noted that all these values must be increased by the transmission propagation delays, which are listed in Table 8.

Table 6: Maximum ISUP Signal Transfer Delays for Each National Component

Country size	Percent of connections	Delay (ms) Mean	95%
Large-size	50%	675-714	<=904-941
	95%	900-952	<=1164-1214
Average-size	50%	450-476	<=637-661
	95%	675-714	<=904-941

Table 7: Maximum ISUP Signal Transfer Delays for International Calls

Country size	Percent of connections	Delay (ms) Mean	95%
Large-size to Large-size	50%	2025-2142	<=2421-2538
	95%	2495-2638	<=2933-3076
Large-size to Average-size	50%	2250-2380	<=2677-2797
	95%	2720-2876	<=3177-3333
Average-size to Average-size	50%	2475-2618	<=2913-3056
	95%	2965-3134	<=3441-3610

Table 8: Calculated Terrestrial Transmission Delays for Various Call Distances (Source: ITU-T Recommendation Q.706, Table 1)

Arc length (km)	Delay terrestrial (ms) Wire	Fiber	Radio
500	2.4	2.5	1.7
1000	4.8	5.0	3.3
2000	9.6	10.0	6.6
5000	24.0	25.0	16.5
10000	48.0	50.0	33.0
15000	72.0	75.0	49.5
17737	85.1	88.7	58.5
20000	96.0	100.0	66.0
25000	120.0	125.0	82.5

6. Basic Call Flow and Call Setup Delays

The following figure illustrates the simplest call flow for call setup in an ISDN-SS7 environment. The end user terminals are assumed to be ISDN phones and use Q.931 messages (i.e., Setup and Alerting). The switches use ISUP messages to establish inter-switch trunks for the subsequent voice communication.

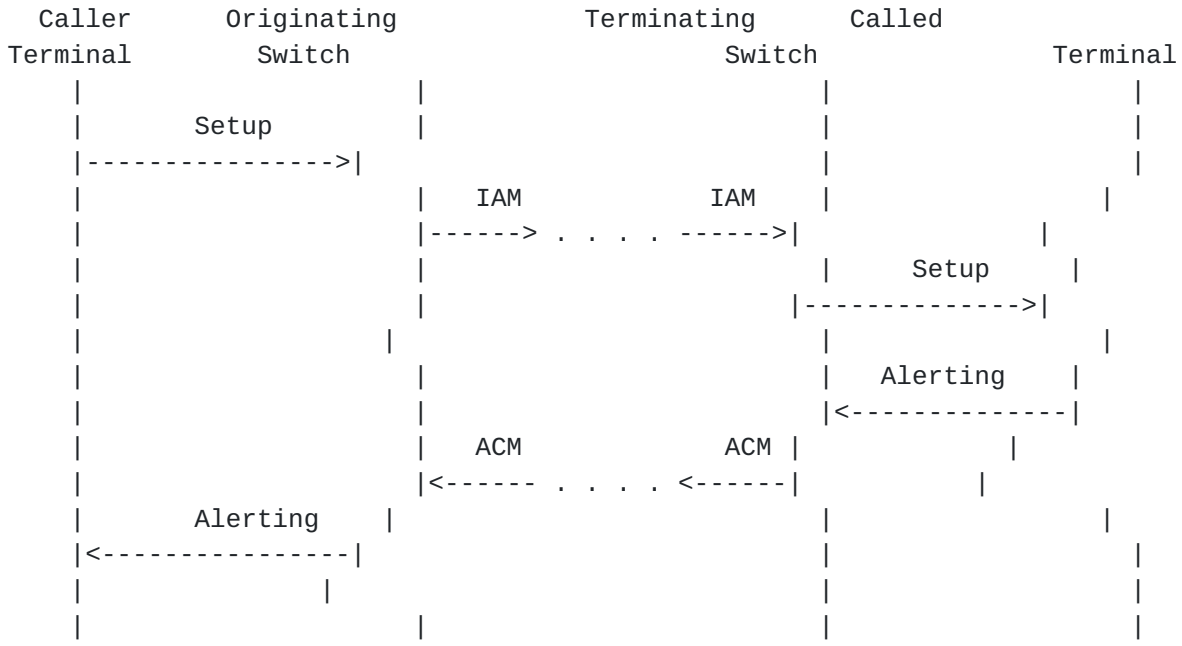


Figure 1: Simple Call Setup Signaling Flow

Using the above call flow, the end-to-end message transfer delays in Tables 6 and 7, and the switch response times for Q.931 messages in Table 3, one can derive the call setup times given in the following tables. Again, all these values must be increased by the transmission propagation delays listed in Table 8.

Table 9: Call Setup Delays for Each National Component

Country size	Percent of connections	Call Setup Delay (ms) Mean	95%
Large-size	50%	2590-2682	<=3007-3099
	95%	3040-3158	<=3497-3615
Average-size	50%	2140-2206	<=2513-2579
	95%	2590-2682	<=3007-3099

Table 10: Call Setup Delays for International Calls

Country size	Percent of connections	Delay (ms) Mean	95%
Large-size to Large-size	50%	5290-5538	<=5909-6157
	95%	6230-6530	<=6903-7203
Large-size to Average-size	50%	5740-6014	<=6387-6661
	95%	6680-7006	<=7378-7704
Average-size to Average-size	50%	6190-6490	<=6863-7163
	95%	7170-7522	<=7893-8245

7. User Expectations

The requirements derived in the previous section should be interpreted as the worst-case requirements. At least in the United States, users of SCN typically experience far less setup delays than the derived delay requirements. With the maturing of Common Channel Signaling (CCS) Network, call setup time has been reduced to a mere one to two seconds [1]. The VoIP networks are expected to achieve the same level of delay

There is no known study on expected setup delays for international calls. As discussed, a HSRC for international working consists of an international component and two national components, and the maximum number of signaling points and STPs in a national component is roughly the same as the number in an international component (Tables 1 and 2). As a consequence, the end-to-end ISUP delays in an international call are roughly three times of those in a national call. On the other hand, the Q.931 signals occur only at the two ends for both national and international calls. Based on these observations, one may expect 2.5-5 second call setup delays to be reasonable for international calls.

8. Post Dial Delay in VoIP Systems

After deriving the PDD requirements in the PSTN, it's important to check if VoIP systems can meet those requirements.

In the following sections, we will evaluate several VoIP systems, illustrate the call flow that contributes to the PDD, and analyze the PDD in terms of delay in network elements. We emphasize that the intend for this analysis is not to set the requirement for VoIP systems, but rather to gain further understanding of the PDD expectation in VoIP services.

For definition of Media Gateway (MG), Media Gateway Controller (MGC), Residential Gateway (RGW), Trunking Gateway (TGW), Access Gateway (AGW), and Signaling Gateway (SG), please refer to [7].

8.1. Methodology and Assumptions

The VoIP systems we intend to investigate are based on the architecture and protocol defined in megaco WG [7]. The set of commands megaco protocol specifies is "Add", "Modify", "Subtract", "Move", "AuditValue", "AuditCapacity", "Notify", and "ServiceChange". The Post Dial Delay is a function of the Response Time in each network element (e.g. RGW, MGC), and the Transmission Delay between network elements. The Response Time can be divided into the Processing Time and the Link Output time. The Processing Time required by a network element depends on the command it receives and the state of the call connection at that time. It is defined as the period that starting when a stimulus occurs at the network element (e.g., when the network element receives the last bit of a message from the incoming signaling link) and ending when the network element places the last bit of the message in the output signaling link controller buffer. [3]

The PDD analysis is based on the call flow derived from the megaco protocol and only the portion that contributes to the PDD needs to be considered. In the following section, we will show only "Add", "Modify", "Notify" and their Reply messages are used to calculate the PDD.

For the purpose of comparing the performance of the VoIP systems with that of the PSTN, we assume network elements in each system have comparable Processing Time for executing the same or similar function in a call connection setup. The comparison then can be made based on the system complexity (i.e. number of components) and the set of messages (i.e. the number and types of commands) need to be exchanged and executed. More specifically, we assume the response time to create a connection in a VoIP system (i.e. Add Termination processing in both the MGC and a MG) is comparable with that of a Connection call segment in a PSTN switch (i.e. the Connection in Table 3, which has a mean value of 300 ms and 95%

of prob. not exceeding 432 ms). The split of this Connection Response Time into delay components in MGC and MG depends on the implementation itself. We use 80-90% of it for MGC processing assuming most of the intelligent functionality of a switch now reside in the MGC. The processing for a Modify Termination command is expected to be close to or less than that of an Add Termination command. Also, the processing time for the TGW is expected to be larger than that for RGW.

Therefore, we define:

* MGC Connection Response Time - The call segment for which MGC processing involves the MGC sending an "Add" command as a result of a stimulus.
* MG Add Termination Response Time - The call segment for which MG processing involves the MG adding a termination to a context and sending a reply message as a result of receiving an "Add" command.
* MG Modify Termination Response Time - The call segment for which MG processing involves the MG modifying a termination in a context and sending a reply message as a result of receiving a "Modify" command.

The Signaling Gateway can reside close to the MGC if not in the same host, the Transmission Delay between them is negligible in comparison with the expected PDD values in Table 9. The Signaling Gateway relays signaling message between the PSTN and the MGC, it is assumed to act like an STP in the PSTN, and the cross-STP delay is used for the Processing Time in the SGs.

Therefore, we define:

* SG Response Time - The call segment for which SG processing involves a SG sending a message to the MGC or the PSTN as a result of a stimulus.

In the VoIP scenario, the processing of the call setup messages in the MGC and AGWs is taken to be comparable to the processing of the ISDN Access message call segment, i.e. Setup message, that contributes to the PDD in the PSTN scenario. Further, the terminating switch in the PSTN usually needs to generate the ringback tone after receiving an Alerting message from the called ISDN terminal. However, in the VoIP systems, the terminating AGW do not have to generate the ringback tone. Instead, the ringback tone can be generated by the originating AGW. Therefore, the processing delay for Alerting message at the terminating AGW in the VoIP scenario can be reduced.

The Transmission Delay in an IP network has different characteristics from that in an SS7 network. We gathered some experiment data from the Internet and applied them for the purpose of this analysis. More data based on the methodology being defined in the IPPM WG can refine the characteristics of the Transmission Delay in the future.

For ease of manipulation, we define:
* Tgc - MGC Connection Response Time.
* Tta - TGW Add Termination Response Time.
* Tra - RGW Add Termination Response Time.
* Taa - AGW Add Termination Response Time.
* Ttm - TGW Modify Termination Response Time.
* Trm - RGW Modify Termination Response Time.
* Tam - AGW Modify Termination Response Time.
* Tcc - Transmission delay between two MGCs.
* Tcr - Transmission delay between a MGC and a RGW.
* Tct - Transmission delay between a MGC and a TGW.
* Tca - Transmission delay between a MGC and a AGW.
* Tia - Transmission delay between a user's ISDN terminal and a AGW.
* Ts - SG Response Time.
* Tisup - ISUP message call segment Response Time.

We assume the delays in network elements are mutually independent of each other and have Normal distributions.

In summary, the tentative statistics we use for this draft is as follows:

Table 11: Network Element Processing Time in VoIP.

	Processing Time (ms)	
	50%	95%
Tgc	255	380
Tra/Trm	30	60
Taa/Tam	30	60
Tta/Ttm	60	120
Tcc	100	200
Tcr	15	20
Tct	20	30
Tca	15	20
Ts	20	40

9. PDD Analysis for VoIP scenarios

In this section, we derived four call connection scenarios. For each scenario, we first illustrate the portion of the call flow that contribute to the PDD, then we calculate the PDD based on the assumed characteristics mentioned in the last section.

9.1. Scenario 1: Two Residential Gateways under a MGC

We start with a simple scenario where two Residential Gateways (RGW1 & RGW2) and a Media Gateway Controller (MGC) are involved in a call connection as shown in Figure 2. Both of the RGWs are controlled under

the same MGC.

Figure 2. Call Connection Model, Scenario 1.

The call flows we use for the PDD analysis in this draft are derived from those in [8], and we substitute the commands with the corresponding ones specified in the latest draft of megaco protocol [7]. The portion of the call flow that affects the PDD in scenario 1 is illustrated as follows:

Usr	RGW1	MGC	RGW2
Off-hook			
(Dialtone)			
Digits	Notify	->	
	<-	Reply	
	<-	Add	
	Reply	->	
		Add	->
		<-	Reply
	<-	Modify	
	Reply	->	
	<-	Modify	
ringback		(Signal)	

The sequence of messages must be processed successfully before a ringback tone can be posted to the caller is as follows:

* A Notify message is generated with collected digits by RGW1.
* The Notify message is transported to the MGC.
* The Notify message is processed by the MGC, call connection resource is set in the MGC, an Add message is generated by the MGC.
* The Add message is transported to the RGW1.
* The Add message is processed by the RGW1, a connection is made in the RGW1, a Reply message is generated by the RGW1.
* The Reply message is transported to the MGC.
* The Reply message is processed by the MGC, call connection resource is modified in the MGC, another Add message is generated by the MGC.
* The Add message is transported to the RGW2.
* The Add message is processed by the RGW2, a connection is made in the RGW2, a Reply message is generated by the RGW2.

* The Reply message is transported to the MGC.

* The Reply message is processed by the MGC, call connection resource is modified in the MGC, a Modify message is generated by the MGC.
* The Modify message is transported to the RGW1.
* The Modify message is processed by the RGW1, the connection in RGW1 is modified, a Reply message is generated by the RGW1.
* The Reply message is transported to the MGC.
* The Reply message is processed by the MGC, a Modify message is generated by the MGC for signaling a ringback tone request.
* The Modify message is transported to the RGW1.

After applying the statistics in Table 11 on delay components above, the PDD for this scenario is calculated as:

PDD = 2*Tgc + 2*Tra + 8*Tcr + Trm ,

and has the statistic of

* Mean value 720 ms
* 95% prob. not exceeding 909 ms

This scenario can be compared with the case in the PSTN that both the calling and called parties are served by the same Local Exchange. The switch response time can be found in Table 3 as:

* Mean value 150 ms
* 95% prob. not exceeding 282 ms

9.2. Scenario 2: Two RGWs under Two Different MGCs

The difference between this scenario and the previous one is that the second Residential Gateway is controlled by a different MGC, i.e. MGC2. Some extra messages need to be exchanged between the two MGCs to achieve the call connection.

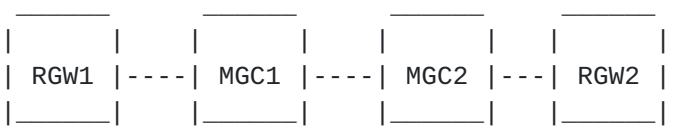
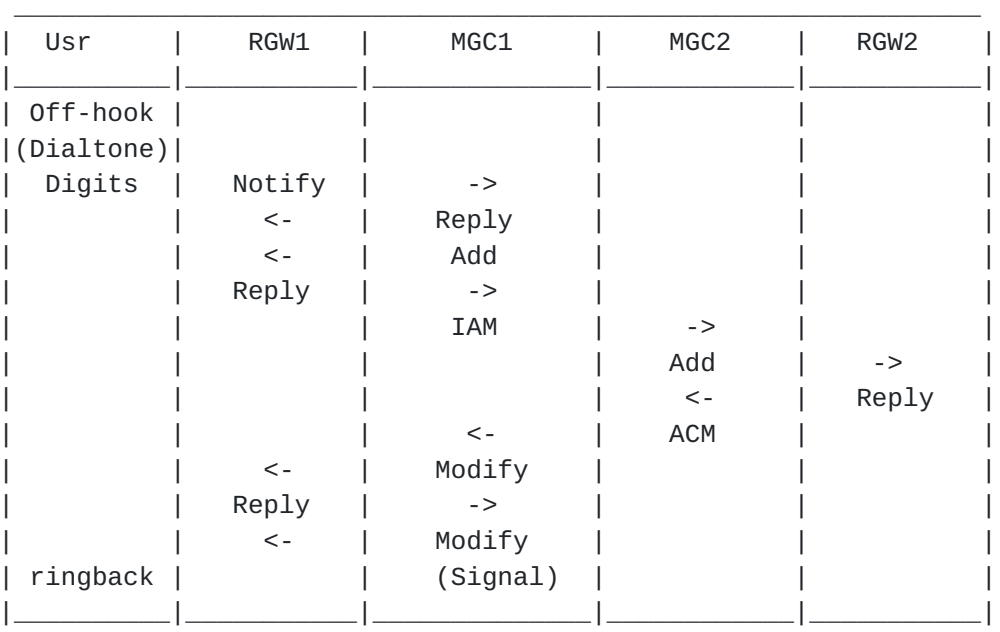


Figure 3. Call Connection Model for Scenario 2.

The portion of the call flow that affects the PDD for this scenario is illustrated as follows:



The additional messages added on top of those in scenario 1 are

```
    ....
* An IAM message is generated by MGC1
* The IAM message is transported to MGC2
    ....
* An ACM message is generated by MGC2
* The ACM message is transported to MGC1
    ....
```

Therefore, by adding the additional two independent random variables to the PDD calculated in section 9.1, the resulting PDD for the current scenario is:

PDD = 2*Tgc + 2*Tra + 8*Tcr + 2*Tcc + Trm ,

and has the statistic of

* Mean value 920 ms
* 95% of prob. not exceeding 1,156 ms

This scenario can be compared with the case in the PSTN that the called party is served by a different Local Exchange than the calling party. Without additional STPs involved in the connection and without counting the transmission delay between two switches, the PDD in the PSTN case is:

* Mean value 904 ms
* 95% of prob. not exceeding 1,127 ms

9.3. Scenario 3: Two ISDN Terminals under Different MGCs

In this scenario, we replace the RGW in the previous section with an Access Gateway, and an ISDN terminal is connected to the Access Gateway. The call connection model is shown in Figure 4.

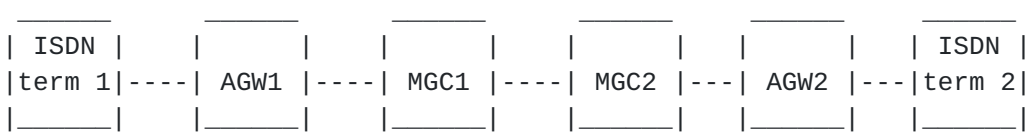


Figure 4. Call Connection Model for Scenario 3.

The portion of call flow that affects the PDD is shown as follows:

Caller	AGW1	MGC1	MGC2	AGW2	Callee
Setup	->-	->			
	<-	Add			
	Reply	->			
		IAM	->		
			Add	->	
			<-	Reply	
			Setup	->-	->
			<-	-<-	Alerting
		<-	ACM		
	<-	Modify			
	Reply	->			
<-	-<-	Alerting			

The ISDN Setup and Alerting messages are exchanged between the ISDN terminal and the MGC via a relay in the AGW using a signaling back-haul protocol. The AGW does not process the message itself.

Note that, after sending out the Setup message to the Called party, the MGC2 can send a provisional message back to the MGC1 to inform it the RTP connection information of AGW2, etc. In this case, the Modify message MGC1 sends to AGW1 can overlay with the Alerting and ACM messages, and thus the PDD can be reduced.

The resulting PDD for this scenario can be calculated as:

PDD = 2*Tgc + 2*Taa + 8*Tca + 2*Tcc + 4*Tia ,

and has the statistic of

* mean value	906 ms
* 95% of prob. not exceeding	1,140 ms

The processing of the ISDN Access message call segment, i.e. Setup message, that contributes to the PDD in the PSTN scenario is replaced by the Call Connection processing delay in the MGC and the Add Termination processing delay in the Access Gateway. Therefore, there is no need to add additional delay to the PDD. And since the PDD does not include the seizure of a ringing circuit and initialization of the audible ring signal to the caller [3], the PDD is over as soon as the caller's ISDN terminal receives the Alerting message. As a result of the functional differences of the network elements between the VoIP systems and the PSTN, the PDD calculated for VoIP in this scenario is better than those in the PSTN that is shown in Table 9.

9.4. Scenario 4: PSTN users connecting to TGWs

The last scenario we analyzed involves two PSTN users connected by two Trunking Gateways under two different MGCs. The call connection model is shown in Figure 5.

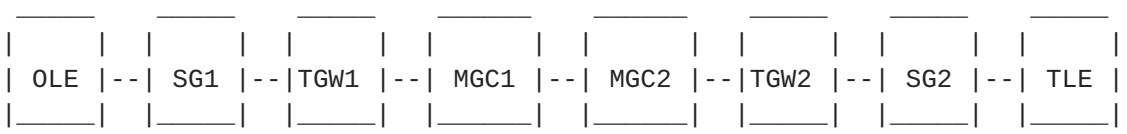
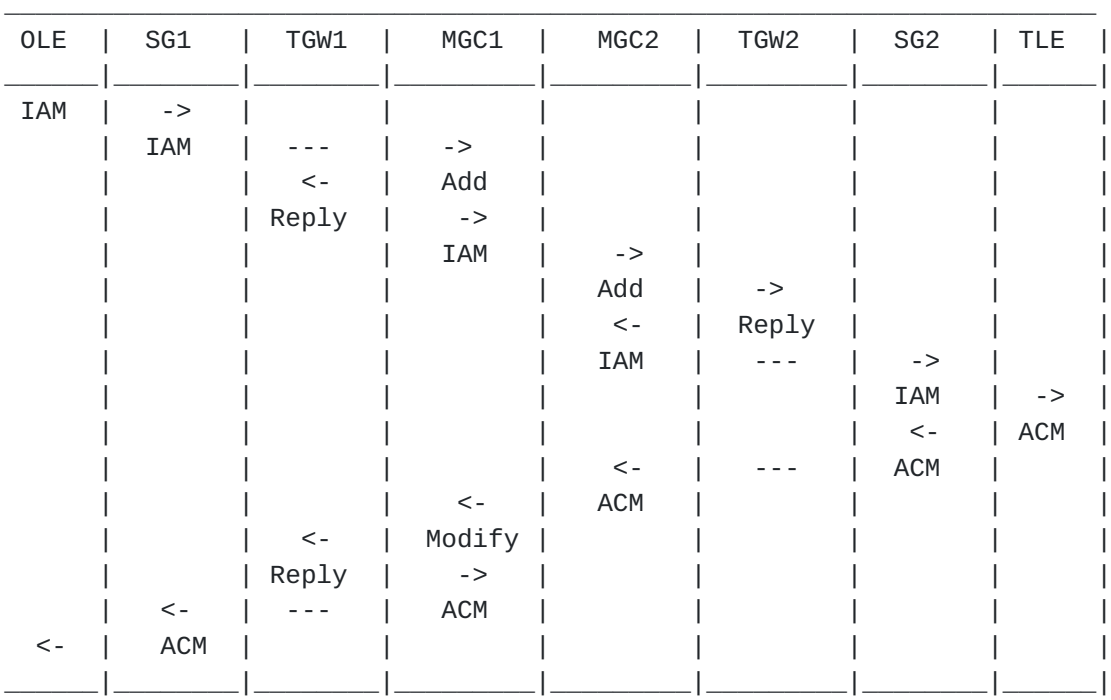


Figure 5. Call Connection Model for Scenario 4.

The portion of call flow that affects the PDD is illustrated as follows:



After the dialed digits are received by the Originating switch in the Local Exchange, they are processed and an IAM message is generated. The timing requirement for this is shown in Table 3. The IAM message is transported to the MGC1 via a relay by a Signaling Gateway (SG1). As mentioned in section 9.1, we use the cross-STP delay of 20 ms to benchmark the performance requirement of the SGs. The same criterion is applied to the ACM message generated by the Terminating switch.

Therefore, the PDD for this scenario is calculated as:

PDD = 2*Tgc + 2*Tta + 4*Tct + 2*Tcc + 4*Ts + 3*Tisup ,

and has a statistics of

* Mean value 1,626 ms
* 95% of prob. not exceeding 1,963 ms

This scenario can also be compared with the case in the PSTN that the called party is served by a different Local Exchange than the calling party. If we assume there are 3 switches and 3 STPs involved in the connection, then the PDD (without counting transmission delay) can be calculated as:

* Mean value 1,295 ms
* 95% prob. not exceeding 1,620 ms

10. Summary of PDD analysis for VoIP systems

As summarized in Table 11, the PDDs for various VoIP scenarios we analyzed are mostly comparable with those in the PSTN. The only exception is scenario 1 where the calling and called parties are in the same Local Exchange. However, the PDD for this scenario is less than 1 second which can meet user's expectation easily. Since the most expensive delay component is the MGC Connection Response Time based on the analysis we have shown, an improvement in this element can bring the PDD performance of VoIP systems closer to if not better than that of the PSTN.

Table 11. Summary of PDD for Various Scenarios.

Scenario	Post Dial Delay in VoIP (ms)		comparable case in PSTN (ms)	
	50%	95%	50%	95%
1. RGW1-MGC-RGW2	720	909	150	282
2. RGW1-MGC1-MGC2-RGW2	920	1,156	904	1,127
3. AGW1-MGC1-MGC2-AGW2	906	1,140	2,140	2,513
4. OLE-SG1-MGC1-MGC2-SG2-TLE	1,626	1,936	1,295	1,620

Acknowledgements

The authors would like to express their gratitude to Dr. Daniel Luan of AT&T Labs for his insight into network operation and valuable suggestions for calculating end-to-end signaling delays as well as call setup delays in section 7.

References

[1] AT&T Webpage, www.att.com/technology/technologists/fellows/lawser.html.

[2] ITU-T Recommendation Q.709, Specifications of Signaling System No. 7--Hypothetical Signaling Reference Connection, March 1993.

[3] Telcordia Technologies Generic Requirements GR-1364-CORE, Issue 1, LSSGR: Switch Processing Time Generic Requirements Section 5.6, June 1995.

[4] ITU-T Recommendation Q.706, Specifications of Signaling System No. 7--Message Transfer Part Signaling Performance, March 1993.

[5] ITU-T Recommendation Q.706, Specifications of Signaling System No. 7--Signaling performance in the Telephone Application, March 1993.

[6] Telcordia Technologies TR-TSY-000511, LSSGR: Service Standards, Section 11, Issue 2, July 1987.

[7] Brian Rosen, et. al., "Megaco Protocol", draft-ietf-megaco-protocol-04.txt, September 21, 1999.

[8] Christian Huitema, et.al., "Media Gateway Control Protocol (MGCP) Call Flows", draft-huitema-megaco-mgcp-flows-01.txt, January 20, 1999.

Authors' addresses

Huai-An Lin
Telcordia Technologies
445 South Street, MCC-1A216R
Morristown, NJ 07960-6438
Phone: 973 829-2412
Email: hlin@research.telcordia.com

Kun-Min Yang
Telcordia Technologies
331 Newman Springs Road, NVC-3X311
Red Bank, NJ 07701
Phone: 732 758-4034
Email: dyang@research.telcordia.com

Taruni Seth
Telcordia Technologies
445 South Street, MCC-1G209R
Morristown, NJ 07960-6438
Phone: 973 829-4046
Email: taruni@research.telcordia.com

Christian Huitema
Telcordia Technologies
445 South Street, MCC-1J244B
Morristown, NJ 07960-6438
Phone: 973 829-4266
Email: huitema@research.telcordia.com