Signaling Bandwidths:
Estimating IP Bandwidths for ECLIPS Signaling Connections

ABSTRACT

The customer’s converged voice & data network carries the signaling for ECLIPS products. These signaling data streams must share the available bandwidth with the customer’s data and telephony media. A robust and responsive communications system is dependent upon adequate bandwidth being available for this signaling. This document examines the characteristics of signaling streams and provides signaling bandwidth requirements for ECLIPS connections and guidelines for engineering the network for specific configurations.
Table of Contents

1. Introduction ........................................................................................................... 3
   1.1. References .................................................................................................. 3
   1.2. Acronyms ................................................................................................... 4
2. Factors influencing signaling traffic ................................................................. 5
   2.1. Link Types ................................................................................................. 5
   2.2. Traffic Estimation ...................................................................................... 6
3. Procedure for Estimating Signaling Bandwidths ............................................. 7
   3.1. Approximations ......................................................................................... 7
   3.2. Estimated IP call Signaling Bandwidths (why averages don’t work) .... 8
   3.3. Link Estimation Procedure ....................................................................... 8
4. Cautions and Limits ............................................................................................ 12
5. Conclusions ......................................................................................................... 13
Appendix A. Detailed Link values .......................................................................... 15
Appendix B. Background on the requirements and model. ............................... 17
1. Introduction

As telephony is migrated to Voice over IP (VoIP), voice samples and signaling must traverse the customer’s network. The bandwidth required to provide adequate service must be provided. Availability of bandwidth affects the economics of the configuration. For example: Is the currently available bandwidth adequate or does more bandwidth need to be purchased? Additional bandwidth will have costs associated with it, and lead time for ordering bandwidth can be long, so this will be a factor in when an implementation can be cut. Providing too much bandwidth can be expensive; not providing adequate bandwidth can have disastrous impacts on the performance of the system.

The number of users, the number of calls, the type of signaling links, and the transport mechanism will all affect the bandwidth required. In this document specific numbers are given for several links. These numbers can be used to estimate bandwidth needs. Reasonable approximations are suggested and several examples are provided. These numbers represent typical calls so understanding the specific implementations and applying them appropriately is important.

For the discussions in the document, a system consists of one or more media controllers, one or more gateways and numerous endpoints. These elements are grouped at nodes or regions which are connected by links that are either limited in bandwidth, expensive or both. See Figure 1 as an example.

1.1. References


[3] Bandwidth estimating calculators are being developed by Avaya Labs and will be available in the future through support.avaya.com

1.2. Acronyms

BH    Busy Hour
BHC   Busy Hour Calls
CoS   Class of Service:
G600  Avaya™ G600 Media Gateway: This model is connected via the IPSI (IP Server Interface)
G700  Avaya™ G700 Media Gateway. This model is connected to a CLAN via an H.248 link
GoS   Grade of Service: the probability that a user will not receive a response within the required time
Hold time In traffic engineering this refers to the length of an active call.
IPSI  IP Server Interface. A Circuit pack: TN2312 which carries signaling for the Avaya G600 Media Gateway
MC    Media Controller: either a CLAN (TN799) or an Avaya™ S8300 Media Server
PSTN  Public Switched Telephone Network
QoS   Quality of Service.
S8300 Avaya MultiVantage™ Software running on an Avaya™ S8300 Media Server: A media server that is provided as a blade which fits in the Avaya G700 Media Gateway and can control several G700 Media Gateways or act as a Local Survivable Processor (LSP)
S8700 Avaya MultiVantage Software on an Avaya™ S8700 Media Server: A media server pair, which can control a large system through traditional port networks, Avaya G600 and G700 Gateways.
VoIP  Voice over IP.
2. Factors influencing signaling traffic

The link type, call type, the number of users and the number of call attempts will all affect the signaling traffic across the WAN.

The incremental call traffic for IP Phones, G600s and G700s is very similar in volume; however, the G600 has a higher base level of traffic. Both gateways have periodic maintenance related traffic. If very narrow links are used, that is links with high utilization, then latency and lost packets will severely impact the performance. This will result in slow response and resetting of the gateway, causing periods of lost service and dropped calls.

The type of station making the call will also influence the amount of signaling traffic. Values are given in the Appendix at the end of this document, with a typical value and a range. Estimates should be increased if a large number of display or lamp updates are expected.

Knowing the call types, number of users and the number of call attempts will enable the appropriate calculation of bandwidth needs. The hold time of each call is not a factor for signaling. The amount of signaling traffic is independent of the length of the calls, and depends only on the number of calls attempted. Note, if signaling and bearer traffic are being routed across the same link, hold time will have a large impact on the bearer traffic, and hence the overall network usage.

2.1. Link Types.

Three types of links are relevant and each has different characteristics. The types are Individual Set Signaling, G700 - H.248 Signaling and G600 - IP-Server Signaling. The characteristics are summarized here; the detail is included in the Appendix.

**Individual Set Signaling:** This is the link between and individual IP phone and a CLAN board or an S8300. It includes UDP messages used for Gatekeeper registration, and a TCP session that provides the signaling connection while the set is registered. The traffic from each set is nearly independent from other sets. It includes both TCP and application layer keep-alives which are nearly constant. The variable element is the signaling to initiate and terminate calls, as well as display and lamp updates. This data traffic is dependant primarily on the number of calls, and to a much lesser degree the type of calls. When calculating signaling bandwidth, it is important to determine the location of the CLAN to which the phones register.
**G700 - H.248 Signaling:** When a remote office is served by a G700 Gateway, the signaling for the gateway is carried over a TCP session carrying H.248 signaling traffic. Gateway control is the major contributor for this link. There is also periodic background maintenance and application layer and TCP keep-alives. If an LSP is associated with the G700 there will be low-level keep-alive activity and occasional translation downloads. Incremental traffic will be associated with G700 endpoints and control of the VoIP resource. IP telephones associated with the Gateway will each have their own signaling link to the media controller. Loss of this link will cause the gateway to seek an alternate Gatekeeper, disrupting service.

**G600, IP-Server Signaling:** A G600 gateway can be remoted, providing rich endpoint support at the remote location. This type of gateway uses a proprietary protocol well suited for traditional endpoints. Constant background sanity scanning and keep-alive is performed. Loss of this link will cause the gateway to reset. Call control and periodic maintenance bandwidth is added to this and depends on the call traffic and the size of the gateway. Loss of link will cause reset the gateway and cause loss of service [2].

These characteristics will be used to determine the bandwidth needed for each element of the network.

### 2.2. Traffic Estimation

The number of call attempts is a critical factor in estimating the required signaling bandwidths. Where legacy systems are being replaced and existing call records can be used, these will provide the best estimates. If records do not exist some common engineering guidelines are provided in Table 1. The Avaya design center can provide additional guidelines.

<table>
<thead>
<tr>
<th>Call Type</th>
<th>Calls/Endpoint/Hour</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Business</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Light</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>General Business</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moderate</td>
<td>2</td>
<td>Rough guidelines for office workers, Use cautiously and only if actual numbers are unavailable.</td>
</tr>
<tr>
<td>General Business</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heavy</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Call Center</td>
<td>10-12</td>
<td>Always perform final call center estimates input from customer configuration.</td>
</tr>
</tbody>
</table>

Table 1 Call Capacity Guidelines
The values presented in Table 1 are estimates and guidelines. They rely on a large number (dozens to hundreds) of endpoints for validity. They also assume that there will be a mix of active phones as well as conference room, lobby, and vacant desks where few or no calls will be made. If growth is anticipated, include it in the bandwidth calculations, even if endpoints will not be provisioned immediately. The type of endpoint, type of call and feature assignments to stations will all affect the call traffic, but only second in degree to the number of attempts. Special care must be taken in engineering Call Center and ACD applications.

3. Procedure for Estimating Signaling Bandwidths

Carefully describe the topology of the network to be implemented. This should include the proposed signaling paths and the bandwidth available.

- For each node, identify the type of gateway, if applicable. The number of users, type of sets, the number of trunks, the busy hour (BH) and the predicted busy hour calls (BHC).
- Use the values provided in the tables in the appendix or the approximations in the next section to calculate the IP traffic for each link.
- Apply the appropriate factors for transport overheads.
- Calculate the bandwidth needed to minimize latency and loss.

At this point the suitability of the existing network to transport the proposed signaling can be assessed, or an appropriate proposal for new service can be made.

3.1. Approximations

Detailed and precise estimations are difficult to make and rarely worth the effort. It is important to quickly and accurately generate guidelines that will meet system performance criteria. Some approximations are in order.

Required Information:
- The network topology
- The number of endpoints
- The estimated call load.

Valid approximations:
- All calls as IP set calls – IP calls generally have the heaviest signaling traffic.
- Assume that originating and terminating (calling and called) traffic is in the same region. This double counts some traffic, but simplifies the calculations.
- Protocol overheads (For PPP and Frame Relay) at 10%
This will yield slightly higher bandwidth requirements than refining the call mix, but not enough to refine the model. These estimates are for fixed rate links that do not share capacity with bearer traffic or non-telephony data.

3.2. Estimated IP call Signaling Bandwidths
(why averages don’t work)

While the average bandwidth required over an hour for completing calls (typically 1-3 per station for general business) is low, sufficient bandwidth must be available at the instant the user wishes to originate a call for the system to perform well. Most notably, with insufficient bandwidth, the delay from off-hook to dial-tone will increase. This delay is already affected by call server load and the network delay. A typical requirement is for the system to provide dial-tone within 350ms of a station going off hook at least 98% of the time. Handset users may have a lower expectation. A grade of service (GoS) can be defined as the probability that a user will not receive a response within the required time. Headset users may have higher expectations. In addition to queuing and serialization delay, there will be carrier or propagation delays. Table 2 is presented to show typical values computed with eight bytes/packet overhead and is used in the examples. More detailed discussion of the requirement and model is in appendix B.

<table>
<thead>
<tr>
<th>Call per hour (up to)</th>
<th>Bandwidth For GoS (kbps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>20.5</td>
</tr>
<tr>
<td>200</td>
<td>21.5</td>
</tr>
<tr>
<td>300</td>
<td>22.5</td>
</tr>
<tr>
<td>400</td>
<td>23.5</td>
</tr>
<tr>
<td>500</td>
<td>24.5</td>
</tr>
<tr>
<td>1000</td>
<td>29</td>
</tr>
<tr>
<td>2000</td>
<td>39</td>
</tr>
</tbody>
</table>

Table 2 Bandwidth required for 350ms Dial-tone 98%

Note: The techniques used to estimate these requirements are approximate for a finite user set. This table provides a good estimate for low delay networks requiring nominal performance (350ms/98%) A calculator is under development and will be available soon [3].

3.3. Link Estimation Procedure.

The best way to communicate the estimation procedure is to show several examples. The configuration used for these exercises has been written down, and the necessary data is in Figure 1and Table 3. This example includes a Media Server (e.g. S8700 pair) with Port Networks at node A; G700s with traditional endpoints, PSTN connections and IP sets at nodes B and E. Nodes C and D include G600 media gateways with traditional endpoints, PSTN connections and IP sets. Note that calculating the bandwidth between nodes A and D will require including the bandwidth required for node E.

Using the example information along with the data in Table 2 and the Appendix, the recommended bandwidth for each link is calculated using the approximations described in section 3.1. The link utilization is calculated to give the reader an indication of utilizations that yield good performance.
Figure 1 Example Network

Table 3 Example Configuration

<table>
<thead>
<tr>
<th>Node</th>
<th>Gateway</th>
<th>PSTN trunks</th>
<th>Traditional Sets</th>
<th>IP Sets</th>
<th>Busy Hour Calls</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>B</td>
<td>3 G700</td>
<td>24</td>
<td>5</td>
<td>90</td>
<td>200</td>
</tr>
<tr>
<td>C</td>
<td>1 G600</td>
<td>32</td>
<td>100</td>
<td>100</td>
<td>300</td>
</tr>
<tr>
<td>D</td>
<td>1 G600</td>
<td>24</td>
<td>50</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>E</td>
<td>1 G700</td>
<td>16</td>
<td>5</td>
<td>70</td>
<td>250</td>
</tr>
<tr>
<td>F</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>40</td>
<td>120</td>
</tr>
</tbody>
</table>
For the following examples some key numbers have been pulled out of Appendix A and are used along with the information in Table 2 and Table 3. PPP overheads have been added to simplify calculations in the examples.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Abbreviation</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Idle bandwidth for a set</td>
<td>SetIdleBW</td>
<td>55bps</td>
</tr>
<tr>
<td>Idle bandwidth for a G600 gateway</td>
<td>G600IdleBW</td>
<td>11kbps</td>
</tr>
<tr>
<td>Idle bandwidth for a G700 gateway</td>
<td>G700IdleBW</td>
<td>55bps</td>
</tr>
<tr>
<td>Bandwidth required to achieve the specified Grade of Service</td>
<td>GoSBW</td>
<td>See Table 2</td>
</tr>
<tr>
<td>Signaling per Call</td>
<td>SigCall</td>
<td>4000 Octets</td>
</tr>
</tbody>
</table>

Table 4 Abbreviations and values used in examples

**Example 1: Node B to Node A. (G700)**

This location has three G700 Gateways serving 48 PSTN connections, five traditional phones and 90 IP stations. The location operates as an independent branch, with no IP calls going to other regions. The G700 and the IP phones are all place in the same network region, with no codecs administered for inter-region calls. In this case the required bandwidth will be the sum of the idle bandwidth for sets and gateway plus the call load bandwidth.

Required bandwidth:

\[
3 \times G700\text{IdleBW} + 90 \times \text{SetIdleBW} + \text{GoSBW} \\
3 \times 55\text{bps} + 90 \times 55\text{bps} + 21.5\text{kbps} = 26,615\text{bps}
\]

In this case, a 32kbps links, or a 32kbps priority queue will provide the necessary performance.

Note that the required GoS BW from Table 2 is higher than the actual signaling traffic. Given a 32kbps link and using the 4000 octets per call to compute the actual load generated by signaling.

The actual bandwidth used will be:

\[
(3 \times G700\text{IdleBW} + 90 \times \text{SetIdleBW} + (\text{BHC} \times \text{SigCall} \times 8))/3600 = 6893\text{bps}
\]

Yielding a utilization of

\[
6893/32,000 = 22\%
\]

**NOTE:** The actual bandwidth must be less than the GoS bandwidth, which should be less than or equal to the provisioned bandwidth, to assure a responsive system.
Example 2: Node C to Node A. (G600)
This location has 32 trunks, 100 traditional sets and 100 IP phones, which are
registered in this port network. This location also operates as an independent
branch, with no IP call going to other regions. The media processor cards, CLANs
and the IP phones are all placed in the same network region, with no codecs
administered for inter-region calls.

Required bandwidth:
\[
1 \times \text{G600IdleBW} + \text{GoSBW} \\
1 \times 11\text{kbps} + 22.5\text{kbps} = 33,500\text{bps}.
\]

In this case a 64kbps or 56kbps link or priority queue will provide the necessary
performance.

The actual bandwidth used will be:
\[
1 \times \text{G600IdleBW} + \left( \frac{\text{BHC} \times 4000 \times 8}{3600} \right) \\
11000 + \left( \frac{300 \times 4000 \times 8}{3600} \right) = 13.7\text{kbps}
\]

Link utilization with a 64kbps link is approximately 21%

The IP set idle bandwidth is not included in this calculation as it travels between the
sets and the CLAN within the node. The keep-alive traffic is included in the
gateway idle bandwidth.

Example 3: Node E to Node D. (G700 behind G600)
This location has a G700 with 16 trunks 5 traditional sets and 70 IP stations. With
250 calls during the busy hour, this location has more calls than a typical business.
Again, for this example, this location operates as an independent branch. The
primary media controller for the G700 is on a CLAN in the G600 at node D.

Required bandwidth:
\[
1 \times \text{G700IdleBW} + 70 \times \text{SetIdleBW} + \text{GoSBW} \\
1 \times 55\text{bps} + 70 \times 55\text{bps} + 22,500\text{bps} = 26,405\text{bps}.
\]

This link could theoretically fit in a 32kbps channel. We observe however that this
link will have multiple hops; so using a full 64k channel is warranted to reduce the
delays.

The actual bandwidth will be
\[
1 \times \text{G700IdleBW} + 70 \times \text{SetIdleBW} + \left( \frac{\text{BHC} \times 4000 \times 8}{3600} \right) \\
1 \times 55 + 70 \times 55 + \left( \frac{250 \times 4000 \times 8}{3600} \right) = 6127\text{bps}
\]

Link utilization on a 64kbps channel is approximately 10%.

Example 4: Node D to Node A.
This is a lightly loaded G600 gateway, however, the data infrastructure routes the signaling traffic from node E through node D. This will allow us to size this link on the aggregate traffic instead of adding the load. The call load will be 350 Calls/hours, assuming that both locations have the same busy hour.

Required Bandwidth:
\[ 1 \times G700IdleBW + 1 \times G600IdleBW + 70 \times SetIdleBW + GoSBW \]
\[ 1 \times 55bps + 1 \times 11kbps + 70 \times 55bps + 23.5kbps = 38405bps. \]

In this case a 64kbps link, or a 64kbps priority queue will provide the necessary performance. The actual bandwidth will be:

\[ 1 \times G700IdleBW + 1 \times G600IdleBW + 70 \times SetIdleBW + (BHC*SigCall*8)/3600 \]
\[ 1 \times 55bps + 1 \times 11kbps + 70 \times 55bps +350*4000*8/3600 = 18.6kbps \]

Link utilization for 64kbps link is approximately 28%

Had the call load bandwidth for each node been estimated independently, then the BW requirement would have exceeded 64k, suggesting that more bandwidth was warranted. This would have been an unnecessary expense.

4. Cautions and Limits

These models work well for a large number of cases, however, each design should be reviewed for unusual cases that might make the general model insufficient. The following points should be reviewed to see if adjustments are necessary should be made in the bandwidth calculations.

This model assumes a 350ms/98% off-hook to dial-tone requirement. This requirement is a compromise between headset and hand set users. Handset users will easily tolerate 500ms of dial-tone delay, where headset users make become annoyed even at 350ms. Some design guidelines suggest 100ms for headset users. Call centers and ACDs tend to be headset operations and as such may require faster links than indicated by the model.

If a large number of lamp or display updates are expected the bandwidth estimates should be increased. The bandwidth from evenly distributed updates can simply be added to required bandwidth. Avoid placing a large number of lamps on endpoints that must be updated simultaneously, for example queue status indicators in a call center.

Very small numbers of sets with low traffic load will still require on the order of 12kbps of available bandwidth. Any reduction will result in dial-tone delays and sluggish response.

The customer’s network most likely has delay in addition to the queuing delays derived here. This model assumed that the network had approximately 40ms of
delay in each direction. If non-queuing delays exceed 40ms, higher speed links will be required to get the overall off-hook to dial-tone delay into the targeted range.

This model is base on a large number of sets and assumes an infinite pool of callers, as call center volumes are approached this assumption is no longer true. This would tend to drive the model to require less than the calculated bandwidth.

If a priority queue provides the signaling bandwidth, then careful attention should be paid to the MTU on the link. A 1500 octet packet at 256kbps will take 47ms to transmit. Even if the signaling is in a strict priority queue a packet arriving just as a large MTU data packet is being transmitted will incur significant delays. Follow the same recommendations as given for bearer traffic in [ 1 ].

Multiple hop systems will incur queuing delays at every router. Engineer the signaling channels generously to avoid adverse effects.

If there is a busy condition that is worse than the normal busy hour, determine the minimum grade of service that is acceptable to provide. Advertising campaigns, market crashes and natural disasters can all cause loads that exceed the busy hour norm. If slow response is acceptable during these events, there is no need to explore further. However, if degradation in service is not tolerable; engineer to these expected levels of traffic.

These estimates are for running steady state systems. Initialization requires assurances in design that TFTP and DHCP servers are properly located within the network to support initialization of the telephones. For example: downloading firmware to a large number of sets in a system where the WAN was not engineered to support bearer, could severely tax the system.

All else being equal, there will be less incremental bandwidth and better response if the phone is registered to a PN at its location, than if it registers to a PN at the home location.

5. Conclusions

These procedures and data can yield useful guidelines for signaling bandwidths links, either on dedicated facilities or in a priority queue. Given the proper input on call traffic and network topology acceptable grade of service can be engineered.

All information in this document is subject to change without notice. Although the information is believed to be accurate, it is provided without guarantee of complete accuracy and without warranty of any kind. It is the user’s responsibility to verify and test all information in this document. Avaya shall not be liable for any adverse outcomes resulting from the application of this document; the user must take full responsibility.
Appendix A. Detailed Link values

The information provided here is the best information at the time this paper was written. It is believed to be accurate and was collected from on pre-release versions of MultiVantage 1.1 and 1.2. All values are one-way bandwidths. All values are IP layer; add the appropriate transport overhead. Incremental traffic includes one origination and one answer. Older dual-connect IP phones are not addressed here.

Individual Set Signaling: This is the link between and individual IP phone and a CLAN board or an S8300. It includes a few UDP messages used for Gatekeeper registration, and a TCP stream that provides the set signaling while the set is registered. This signaling is of concern when the set must communicate to the gatekeeper over a WAN link. In most configurations the bearer traffic traverses the same link and will be the key driver in link size. The important exception for this exercise is an IP phone in a remote location that has a G700 gateway. The phone will register on a CLAN at main location, but may only have bearer traffic within the remote location, with dial-tone and trunks provided from the G700.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Range</th>
<th>Units</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Idle Bandwidth</td>
<td>50</td>
<td>&lt;50</td>
<td>bps</td>
<td>The application keep-alive message is exchanged with the server approximately once per minute. This rate decreases as the number of registered stations increase.</td>
</tr>
<tr>
<td>Incremental traffic per call.</td>
<td>4000</td>
<td>3000-6000</td>
<td>octets</td>
<td>Downlink (from MC to Set). The uplink traffic is somewhat smaller. This includes traffic for both calling and called phones.</td>
</tr>
<tr>
<td></td>
<td>52</td>
<td>30-72</td>
<td>packets</td>
<td>This value is useful in calculating overheads.</td>
</tr>
<tr>
<td>Peak Bandwidth</td>
<td>12</td>
<td></td>
<td>kbps</td>
<td>This spike occurs at call initiation, i.e. drawing dial-tone, over a 350ms period.</td>
</tr>
</tbody>
</table>

Table 5  IP Telephone Signaling
**G700, H.248 Signaling:** When a remote office is served by a G700 Gateway, the signaling for the gateway is carried over an H.248 Link. Gateway control is the major contributor for this link. There is also periodic background maintenance. If an LSP is associated with the G700 there will be low-level keep-alive activity and occasional translation downloads. IP telephones associated with the Gateway will each have their own signaling link to the media controller.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Range</th>
<th>Units</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Idle Bandwidth</td>
<td>50</td>
<td>&lt;50</td>
<td>bps</td>
<td>This is a keep-alive message that is exchanged with the server approximately once per minute. Gateway keep-alive messages are constant.</td>
</tr>
<tr>
<td>Incremental traffic per call.</td>
<td>3000</td>
<td>1200-5000</td>
<td>octets</td>
<td>With both endpoints in the same G700.</td>
</tr>
<tr>
<td></td>
<td>27</td>
<td>13-40</td>
<td>packets</td>
<td></td>
</tr>
<tr>
<td>Peak Bandwidth</td>
<td>NA</td>
<td></td>
<td></td>
<td>The number of endpoints and the traffic load will determine peak.</td>
</tr>
</tbody>
</table>

Table 6 G700 Signalling

**G600, IP-Server Signaling:** A G600 gateway can be remoted, providing rich endpoint support at the remote location. This type of gateway uses a protocol well suited for traditional endpoints. A fairly constant background sanity scanning is performed. Call control and periodic maintenance is added to this.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Range</th>
<th>Units</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Idle Bandwidth</td>
<td>10</td>
<td>6-10</td>
<td>kbps</td>
<td>This background sanity checking remains constant, independent of load. Downlink traffic is greater than uplink. Allow 10kbps for good performance.</td>
</tr>
<tr>
<td>Incremental traffic per call.</td>
<td>2500</td>
<td>1200-5000</td>
<td>Octets</td>
<td>With both endpoints in the same G600.</td>
</tr>
<tr>
<td></td>
<td>21</td>
<td>&lt;40</td>
<td>packets</td>
<td>The number of packets per call will decrease as the traffic increases. Messages will be aggregated into larger packets</td>
</tr>
<tr>
<td>Peak Bandwidth</td>
<td>NA</td>
<td></td>
<td></td>
<td>The number of endpoints and the traffic load will determine peak.</td>
</tr>
</tbody>
</table>

Table 7 G600 Signalling Bandwidth
Appendix B. Background on the requirements and model

This paper uses several assumptions about the required performance of the system and the model used to determine the underlying network requirements. This section provides some background for the interested reader to better understand how these assumptions were derived.

Dial-tone in 350ms

Actually: dial-tone in 350ms or less 98% of the time. This requirement is written with a probability or service level like many telephony requirements. Because engineering to 100% compliance when the call arrival rate is random is very expensive, a few attempts are allowed to miss the requirement, but most will be within the requirement.

The specific requirement used in this document is derived from many years of human factors studies to determine acceptable performance of PBXs. 350ms is a compromise between a handset user’s requirement of 500ms and a headset user’s requirement of 100ms. Two percent is also a common service level, however, the customer can always specify the appropriate level for the business need.

In traditional telephone systems, “registers,” which collect the dialed digits, were a scarce resource. Users were allow to queue up for their availability and dial-tone provide when the resource was available. In the VoIP world, registers are not usually an issue because the digit collection is performed by the set and server exchanging messages. Even though registers are not an issue, time to dial-tone remains a critical performance parameter. With H.323, the call signaling is most intense during the off-hook to dial-tone period. As long as the signaling performs well during this interval, all signaling should be responsive.

Translating Dial-tone delay to Network delay.

Providing “dial-tone” actually involves more than simply switching on a tone. In addition to verifying that the system has the resource to process another call, there is an exchange of several messages with the terminal to set up displays, buttons and lights. Each of these is an message across the network that can be delayed in a router’s queue. (For these discussions, both serialization and internal queueing delay are considered. In queueing theory terms, the time in system is the sum of the time in queue and service time. [4]) Analysis of the set-up sequence shows that six messages must traverse the network for this initial set-up. It is assumed that five of these pass with average delay, but one gets caught in a long queue, delaying the delivery of dial-tone. Obviously, the network delay must be much less than the dial-tone delay. To meet a 350ms delay 98% of the time, the network must be engineered so that the time in system (queue+serialization(service)) is less than 150ms, 98% of the time.
Queueing Model

A basic M/M/1 queue with exponential interarrival times and service times, a single server, infinite queue length and infinite customer population was chosen. This model is mathematically tractable and should provide a reasonable estimate for the problem in question. Inspection of a call trace shows that the service times (packet sizes) approximate an exponential distribution. Exponential arrivals are well established for telephony processes. With a single WAN link, the single server works well. The customer population is limited, but for large numbers of customers (terminals) this is a good approximation, for small numbers of customers it provides an upper bound. The infinite queue assumption is more complicated, as a finite queue with discard will have shorter average queue delay, but messages that are discarded will require retransmission, giving a longer delay for exchanges.

Table 2 Looks like a Straight Line.

Yes, manipulation of the M/M/1 equations yields a straight line to describe the bandwidth needed to meet the desired performance. This function (in units of bits/second) has an intercept:

\[- \text{AvgPktSize} \times 8 \times \ln(\text{AcceptPerform}) / \text{AcceptDelay}\]

Where:

- \text{AvgPktSize} is the size of the average packet
- \text{AcceptDelay} is the design limit on the maximum per packet delay that can be tolerated
- \text{AcceptPerform} is the fraction of time that the acceptable delay can be exceeded

The slope of this line is

\[\text{AveragePktSize} \times \text{AvgPktCall} \times \text{BHC} \times 8/3600\]

Where:

- \text{AvgPktCall} is the average number of packets per call
- \text{BHC} is the number of call in the busy hour.

In other words, the intercept in dependent on the performance level desired and the slope is strictly a function of the actual bandwidth being used.

Using a priority queue

The examples state that using a priority queue is the same as using a dedicated link. A priority queue will usually perform with less average latency than a dedicated link. When the default queue has no traffic in queue the priority queue will be emptied faster and there will also be less serialization delay. The exception to this improvement is when large packets are being transmitted on the default queue and the line speed is relatively low, then completing this large packet may delay servicing the priority queue.