VoIP: How to Plan for the Bandwidth and Calculate the Cost Savings

Bill Treneer, Global Knowledge Instructor, CCSI, SME Cisco Multicasting

Introduction

The economic drivers for voice and data integration using voice over IP (VoIP) are catching the attention of CFOs, CIOs and others involved on the cost side of any business. This white paper will show the cost justification for voice and data network integration and how much increase in bandwidth will be necessary once voice traffic is added to the traditional data traffic being carried across a wide area network (WAN).

Before VoIP, a network for an institution had two distinct infrastructures—one for voice and one for data:
With VoIP, a network for an institution has just one integrated infrastructure:

At first glance, it seems that without the PSTN infrastructure the cost justifications for VoIP are simple. While PSTN usage can be reduced considerably, it cannot be eliminated totally since the PSTN must handle call overflow and any calls not destined to the remote sites. Also, the extra voice over IP traffic may cause an increase in WAN bandwidth. The cost for this extra capacity will have to be subtracted from the savings achieved from reduced PSTN usage. The step-by-step cost justification example below can be used as a model for other network integration projects.

**Step 1**

Obtain a busy hour and an overall traffic study from the voice service provider. Most PBXs provide similar statistics. Use the study to determine the number of hours of voice conversation or Erlangs of voice traffic the new integrated voice/data network is to handle during the busiest hour of the average day of the busiest month.

**Sample Study**

Company XYZ has a Frame-Relay connection in San Jose that is a T-1 physical connection with two 256K CIRs. One 256K CIR is for Houston and the other is for Washington, DC. The average data traffic at San Jose is 150Kbps with a peak of 300Kbps during the busiest hour. XYZ would like to run the voice traffic over the existing frame-relay network if possible with no upgrades of bandwidth, but would upgrade if the costs were justified. The two remote sites have a T-1 physical connection with a 256K CIR. The average data traffic is 75Kbps with a peak of 150Kbps. Company XYZ also has a PBX in San Jose and a PBX in each of its remote sites in
Houston and Washington, DC. San Jose has 14 PSTN lines currently for voice traffic between the PBX and the Central Office (CO) switch of the local telephone company. All sites pay $32 per month for PSTN lines.

Call volume to and from the remote sites for San Jose is 600 calls per day with an average call time of three minutes. Grade of service or Blocking Factor is the percent of calls that are busy during the busiest hour of the organization’s day. XYZ is willing to live with 5% busy grade of service. The busiest hour of the day handles 20% of the traffic. No other hour handles more than 12% of the voice traffic. Each remote site gets about the same number of calls from San Jose. Therefore, call volume to and from headquarters and each site is 300 calls per day with an average call time of three minutes. The remote sites can also live with a 5% busy grade of service and the busiest hour of the day handles 20% of the traffic. Like HQ, no other hour handles more than 12%. The two remote sites have 10 trunk lines each for voice traffic. For this example, assume as a worst case that the busy hour of voice traffic is the same hour as it is for data.

Solution
HQ in San Jose has 600 calls to the two remote sites with each call averaging three minutes. Therefore, the total is 1,800 call minutes or 30 hours of call volume per day.

\[
\text{600 calls} \times 3 \text{ min. per call} = 30 \text{ Hours of traffic} = 30 \text{ Erlangs}
\]

An Erlang is a unit that represents one hour of call volume, so San Jose has 30 hours of call volume to Houston and Washington DC each day. When designing networks, especially those transporting voice, it is best to build the infrastructure for the busiest hour of the day. Therefore, do not assume that the 30 Erlangs are spread out evenly throughout the day. Since 20% of the traffic is handled in the busiest hour, this network will be designed to handle six Erlangs of traffic in a single hour.

\[
(20\%) \times 600 \text{ calls} \times 3 \text{ min. per call} = (20\%) \times 30 \text{ Hours of traffic} = 6 \text{ Erlangs}
\]

Similarly, the remote sites have 15 Erlangs of traffic each and three Erlangs during the 20% busy hour.

\[
(20\%) \times 300 \text{ calls} \times 3 \text{ min. per call} = (20\%) \times 15 \text{ Hours of traffic} = 3 \text{ Erlangs}
\]

Step 2
Use an Erlang table to determine how many PSTN lines can be eliminated once the voice traffic is moved over to the data network.

No telephone system is 100% efficient because a phone call doesn’t come in exactly when a line becomes available. Also, even in the busiest hour there will be minutes when not all the lines are used or some callers get busy signals. Therefore, while one would think that eight phone lines provide eight hours of call volume in an hour, that is only an ideal case that has never been documented. Instead, the Erlang chart uses actual call statistics to give a very accurate estimate of call volume. Again, grade of service or blocking factor is the percent of calls that are busy during the busiest hour of the organization’s day. For instance at a P05 grade of service, eight phone lines provide 4.543 hours of call volume or 4.543 Erlangs. If it is acceptable for 5% of the
Calls to get a busy signal, then a P05 grade of service is used. If 3% is acceptable, then a P03 grade of service is used. If 0.5% is acceptable, then a P005 is used, etc. To transport six Erlangs of call volume at San Jose, the Erlang chart says the equivalent of 10 PSTN lines are needed. To transport three Erlangs of call volume at Houston and Washington, DC, the Erlang chart says the equivalent of seven PSTN lines are needed.

Number of Erlangs increases with the number of simultaneous connections.

Blocking Probability, (Grade of Service), P0X factor: X is a variable

<table>
<thead>
<tr>
<th>Number of Circuits</th>
<th>P003 = 0.003</th>
<th>P005 = 0.005</th>
<th>P01 = 0.01</th>
<th>P03 = 0.03</th>
<th>P05 = 0.05</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.003</td>
<td>0.006</td>
<td>0.011</td>
<td>0.031</td>
<td>0.053</td>
</tr>
<tr>
<td>2</td>
<td>0.081</td>
<td>0.106</td>
<td>0.153</td>
<td>0.282</td>
<td>0.382</td>
</tr>
<tr>
<td>3</td>
<td>0.289</td>
<td>0.349</td>
<td>0.456</td>
<td>0.716</td>
<td>0.9</td>
</tr>
<tr>
<td>4</td>
<td>0.602</td>
<td>0.702</td>
<td>0.87</td>
<td>1.259</td>
<td>1.525</td>
</tr>
<tr>
<td>5</td>
<td>0.996</td>
<td>1.132</td>
<td>1.361</td>
<td>1.876</td>
<td>2.219</td>
</tr>
<tr>
<td>6</td>
<td>1.447</td>
<td>1.822</td>
<td>1.9</td>
<td>2.543</td>
<td>2.961</td>
</tr>
<tr>
<td>7</td>
<td>1.947</td>
<td>2.158</td>
<td>2.501</td>
<td>3.25</td>
<td>3.738</td>
</tr>
<tr>
<td>8</td>
<td>2.484</td>
<td>2.73</td>
<td>3.128</td>
<td>3.987</td>
<td>4.543</td>
</tr>
<tr>
<td>9</td>
<td>3.053</td>
<td>3.333</td>
<td>3.783</td>
<td>4.748</td>
<td>5.371</td>
</tr>
<tr>
<td>10</td>
<td>3.648</td>
<td>3.961</td>
<td>4.462</td>
<td>5.53</td>
<td>6.216</td>
</tr>
<tr>
<td>11</td>
<td>4.25</td>
<td>4.6</td>
<td>5.15</td>
<td>6.3</td>
<td>7.05</td>
</tr>
<tr>
<td>12</td>
<td>4.90</td>
<td>5.25</td>
<td>5.85</td>
<td>7.1</td>
<td>7.95</td>
</tr>
<tr>
<td>13</td>
<td>5.55</td>
<td>5.95</td>
<td>6.60</td>
<td>7.95</td>
<td>8.80</td>
</tr>
<tr>
<td>14</td>
<td>6.20</td>
<td>6.65</td>
<td>7.35</td>
<td>8.8</td>
<td>9.70</td>
</tr>
<tr>
<td>15</td>
<td>6.90</td>
<td>7.35</td>
<td>8.10</td>
<td>9.6</td>
<td>10.60</td>
</tr>
<tr>
<td>16</td>
<td>7.60</td>
<td>8.05</td>
<td>8.85</td>
<td>10.5</td>
<td>11.50</td>
</tr>
<tr>
<td>17</td>
<td>8.30</td>
<td>8.80</td>
<td>9.65</td>
<td>11.35</td>
<td>12.45</td>
</tr>
<tr>
<td>18</td>
<td>9.0</td>
<td>9.55</td>
<td>10.40</td>
<td>12.2</td>
<td>13.35</td>
</tr>
<tr>
<td>19</td>
<td>9.75</td>
<td>10.30</td>
<td>11.20</td>
<td>13.1</td>
<td>14.30</td>
</tr>
<tr>
<td>20</td>
<td>10.45</td>
<td>11.05</td>
<td>12.00</td>
<td>13.95</td>
<td>15.20</td>
</tr>
<tr>
<td>21</td>
<td>11.20</td>
<td>11.85</td>
<td>12.80</td>
<td>14.85</td>
<td>16.15</td>
</tr>
<tr>
<td>22</td>
<td>11.95</td>
<td>12.6</td>
<td>13.65</td>
<td>15.75</td>
<td>17.10</td>
</tr>
<tr>
<td>23</td>
<td>12.70</td>
<td>13.4</td>
<td>14.45</td>
<td>16.65</td>
<td>18.05</td>
</tr>
<tr>
<td>24</td>
<td>13.50</td>
<td>14.2</td>
<td>15.25</td>
<td>17.55</td>
<td>19.00</td>
</tr>
<tr>
<td>25</td>
<td>14.25</td>
<td>14.95</td>
<td>16.10</td>
<td>18.45</td>
<td>19.95</td>
</tr>
</tbody>
</table>
Note: There are several Erlang charts that are used, Erlang B, C, etc. Each one takes into account slightly different parameters such as queuing time, likelihood of busy callbacks, etc. This white paper will use the Erlang B chart. For more detailed information, consult www.erlang.com.

**Step 3**

Calculate the savings from reduced PSTN line and long distance usage.

The number of lines at the HQ site may be reduced from 14 to four since approximately 10 lines worth of traffic are now going across the Frame-Relay network. That is a cost savings of $320/month at HQ. Similar calculations yield approximately six fewer lines at each remote site for a $192/month cost savings per site. Total line savings equal $704/month. Long distance costs of $0.04 per minute for 1,800 total minutes, provide $72.00 of LD savings per day. Assume 21 workdays per month and the monthly LD savings are $1,512. Therefore, total cost savings for this integrated network are: $2,216 per month.

**Step 4**

Determine the type of Codec to use to digitize the voice and decide if compressed Real Time Protocol (RTP) will be used.

Cisco devices use two types of Codec chips that are Digital Signal Processors (DSPs): G.711 and G.729. G.711 digitizes the voice signal at uncompressed 64 Kbps and creates a payload of 160 bytes for the VoIP packet. G.729 uses compression and digitizes the voice signal at eight Kbps and creates a payload of 20 bytes for the VoIP packet. The IP, UDP, and RTP headers are 40 bytes uncompressed, but by using compressed RTP (cRTP), the headers can be reduced to two bytes.

G.711 Codec Digital Signal Processing chips will digitize voice at either 50 packets per second or 33 packets per second. Assume the 50 pps will be used. Therefore, the 64 Kbps calculation is as follows:

\[
\text{160 bytes/packet} \times \text{8 bits/byte} \times \text{50 packets/second} = \text{64,000 bits per second}
\]

However, since the layer 2 frame relay header is four bytes and the layers 3, 4, and 5 are 40 bytes as described above, then the amount of bandwidth needed for G.711 is greater when the headers are included.

\[
\text{204 bytes (headers + payload)} \times \text{64,000 bits per second} = \text{81,600 bits per second}
\]

160 bytes payload only

Therefore, one voice call requires 81.6 Kbps of bandwidth when it becomes a VoIP packet to run on a converged network.

When G.711 is used with cRTP, the calculations change. The layer 2 header is unchanged, but the layers 3, 4, and 5 are 2 bytes instead of 40. One voice call converts to 66.4 kbps with cRTP as shown below.

\[
\text{166 bytes (headers + payload)} \times \text{64,000 bits per second} = \text{66,400 bits per second}
\]

160 bytes payload only
G.729 Codec Digital Signal Processing chips will also digitize voice at either 50 packets per second or 33 packets per second. Assume again the 50 pps will be used. Therefore, the bandwidth calculation with G.729’s 20 byte voice payload is as follows:

\[
20 \text{ bytes} \times 8 \text{ bits} \times 50 \text{ packets} = 8,000 \text{ bits per second}
\]

However, since the layer 2 frame relay header is four bytes and the layers 3, 4, and 5 are 40 bytes as described above, then the amount of bandwidth needed for G.729 is greater when the headers are included.

\[
64 \text{ bytes (headers + payload)} \times 8,000 \text{ bits per second} = 25,600 \text{ bits per second}
\]

Therefore, with G.729, one voice call requires 25.6 kbps of bandwidth when it becomes a VoIP packet to run on a converged network.

When G.729 is used with cRTP, the calculations change once again. The layer 2 header is unchanged, but the layers 3, 4, and 5 are 2 bytes instead of 40. One voice call converts to 10.4 kbps with cRTP as shown below.

\[
26 \text{ bytes (headers + payload)} \times 8,000 \text{ bits per second} = 10,400 \text{ bits per second}
\]

**Step 5**

Calculate the impact on WAN capacity.

HQ bandwidth is 384K CIR with peak traffic of 300K. The remote sites are each 256K CIR with peak traffic of 150K. So the available bandwidth for voice is 84K at HQ in San Jose and 106K at each remote site without going higher than the current CIRs. To be safe, assume again that the busiest hour for voice is the same as the busiest hour for data.

If G.729 is selected with compressed RTP for a Frame-Relay layer 2 header, 10.4 Kbps is needed for each call or line as shown above. Therefore, 10 lines require 104 Kbps extra bandwidth in the busiest hour at HQ and six lines at each remote site require 62.4 Kbps extra bandwidth.

If G.729 is selected without compressed RTP for a Frame-Relay layer 2 header, 25.6 Kbps is selected for each call or line as shown above. Therefore, 10 lines require 256 Kbps extra bandwidth in the busiest hour at HQ and six lines at each remote site require 153.6 Kbps extra bandwidth.

If G.711 is selected with compressed RTP for a Frame-Relay layer 2 header, 66.4 Kbps is needed for each call or line as shown above. Therefore, 10 lines require 664 Kbps extra bandwidth in the busiest hour at HQ and six lines at each remote site require 398.4 Kbps extra bandwidth.

If G.711 is selected without compressed RTP for a Frame-Relay layer 2 header, 81.6 Kbps is selected for each call or line as shown above. Therefore, 10 lines require 816 Kbps extra bandwidth in the busiest hour at HQ and six lines at each remote site require 489.6 Kbps extra bandwidth.
Step 6
Choose the best Codec based on quality versus impact on WAN capacity.

Quality of a voice call is generally judged with the Mean Opinion Score (MOS). The MOS is a score from 1 to 5 with 5 as the highest. Selected listeners are asked to judge the quality of different voice digitizing methods and the following is how the different codecs were rated.

<table>
<thead>
<tr>
<th>ITU Standard</th>
<th>Data Rate</th>
<th>MOS Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCM</td>
<td>G.711</td>
<td>64 Kbps</td>
</tr>
<tr>
<td>ADPCM</td>
<td>G.726 / G.727</td>
<td>16 / 24 / 32 / 40 Kbps</td>
</tr>
<tr>
<td>LDCELP</td>
<td>G.728</td>
<td>16 Kbps</td>
</tr>
<tr>
<td>CS-ACELP</td>
<td>G.729</td>
<td>8 Kbps</td>
</tr>
<tr>
<td>ACELP / MP-MLQ</td>
<td>G.723.1</td>
<td>6.3 / 5.3 Kbps</td>
</tr>
</tbody>
</table>

As is shown, G.729 is quite close to G.711 so only a small loss of quality will occur. G.711 is considered toll quality conforming to accepted telecommunications specifications. Compressed RTP (cRTP) will cause a slight loss of quality that the user will have to judge. Another option to reduce bandwidth even further is to use Voice Activation Detection (VAD). VAD will eliminate silence in voice conversations at the sending end before it is transported across the WAN. Silence can be added back at the receiving end. VAD is more bandwidth efficient for the WAN, but it will further lower the MOS. VAD and cRTP may not provide high quality on all networks. VAD is included in certain models of G.729 codecs.

As shown in step 5, if G.729 is selected with compressed RTP, 10.4 Kbps is needed for each call or line as shown above. Therefore, 10 lines require 104 Kbps extra bandwidth in the busiest hour at HQ and six lines at each remote site require 62.4 Kbps extra bandwidth. Bandwidth at HQ and the remote sites is sufficient and no increase is needed. All of the $2,216 of savings could go to the bottom line or pay a quick return on investment of any needed equipment purchases.

If G.729 is selected without compressed RTP for a Frame-Relay layer 2 header, 25.6 Kbps is needed for each call or line as shown above. Therefore, 10 lines require 256 Kbps extra bandwidth in the busiest hour at HQ and six lines at each remote site require 153.6 Kbps extra bandwidth. Bumping the two CIRs at HQ to 384K each would be required and that would bump each remote site to a 384K CIR. The extra cost for the higher CIRs will be $100 to $400 per month that is easily covered by the $2,216 of savings.
Conclusion

The cost justifications for a VoIP integrated network are simple to calculate and usually large enough to warrant an upgrade in equipment costs if needed. Depending on design, the extra impact on WAN bandwidth capacity may also be minimal.

Learn More

Learn more about how you can improve productivity, enhance efficiency, and sharpen your competitive edge. Check out the following Global Knowledge courses:
- CCDA (Cisco Certified Design Associate) Boot Camp
- IPTD (IP Telephony Design)
- CVoice – Cisco Voice over IP
- CIPT1 – Cisco IP Telephony Part 1
- CIPT2 – Cisco IP Telephony Part 2

For more information or to register, visit www.globalknowledge.com or call 1-800-COURSES to speak with a sales representative.

Our courses and enhanced, hands-on labs offer practical skills and tips that you can immediately put to use. Our expert instructors draw upon their experiences to help you understand key concepts and how to apply them to your specific work situation. Choose from our more than 700 courses, delivered through Classrooms, e-Learning, and On-site sessions, to meet your IT and management training needs.

About the Author

Bill Treneer is a Cisco Certified Systems Instructor (CCSI) and has been contracted by Global Knowledge since 1997. He teaches the entire CCNP and CCDP line of certification courses and is a subject matter expert (SME) in Cisco Multicasting. He has a BS in Electrical Engineering from Washington State University and has been working in the telecommunications and computer industry since 1987.

Sources

www.erlang.com
www.cisco.com