IPCOMM2006

September 25-27 • Gaylord Opryland • Nashville, TN

# How Well Can You Hear Me Now? *Measuring Good Call Quality*

**Dave Chapman**

**President**

**Chapman Consultants**

IPCOMM2006
September 25-27 • Gaylord Opryland • Nashville, TN

# What Will Be Covered

- VoIP vs other applications
- Understanding and detecting typical network impairments in the enterprise network, and their impact on VoIP
- VoIP Quality monitoring
- Data network techniques vs. VoIP requirements
- Wrap up

# General Observations

- VoIP Implementers fall into two broad categories:
- Group A
  - VoIP as a cost-effective alternative to TDM
  - Hybrid PBX implementations (at least to start)
  - VoIP should run itself
- Group B
  - VoIP as a launch pad for IP communications
  - True leap to pure IP PBX
  - VoIP is the most important and complex application, it better be managed carefully
  - VoIP in a converged environment
- For the most part, if you are in group “A” once, you are in group “A” all the way down!

# VoIP vs Other Applications

- VoIP utilizes the network differently than almost any other application
  - Time critical, with near real-time performance requirements
    - Delays greater than 150ms are noticeable by almost every user from the CEO to the janitor
  - UDP (user datagram protocol) vs TCP (transmission control protocol)
    - Somewhat unique among popular network based applications (with the notable exception of network management tools!)

# VoIP vs Other Applications (2)

- Often highly distributed, with components going in and out of use during a “session” (phone call)
  - IP PBX to phones to set up calls, phone to phone, or gateway to phone, then IP PBX at call tear down
- Multitude of “standards” for communications
  - SIP, H323, SCCP, etc (just for call signaling!)
  - G.729, G.711, G.729a, etc (just for carrying payloads!)
  - You get the point

IPCOMM2006
September 25-27 • Gaylord Opryland • Nashville, TN

# Comparison of VoIP to other Apps

| | HTTP | Email | SAP | VoIP |
|---|---|---|---|---|
| Time | Seconds to min. | Minutes and up | Seconds | <150 ms |
| Packet management | Retry, retry retry, until I have them all | Retry, retry, retry until I have them all | Retries!!! | Ignore packets that are late, they are now useless |
| Tolerance for downtime | Let me try again. | Was I expecting mail? | High frustration, business at risk | Life shattering, and possibly life at risk |
| Organizational impact | Large portion, with few “critical users” | Large portion, with many “critical users” | Small portion, but almost all critical | Everyone, including external customers |
| Complexity | Low | Low | High | Very High |
| Support staff | Small | Small | Mid-sized, but very focused | Cross organizational |

**VoIP is more complicated, harder to manage, and has the largest impact on the organization.**

IPCOMM2006
September 25-27 • Gaylord Opryland • Nashville, TN

# One more point…

- Poll:
  - How many people had a phone on their desk
    - 15 years ago?
    - 20 years ago?
    - 30 years ago?
  - Did it work as reliably then as it does now?
  - How many people would say SAP, Email or the Web applications work AS reliably?
  - Of the 4 applications, which could you do without for 4 hours?

IPCOMM2006
September 25-27 • Gaylord Opryland • Nashville, TN

# What Will Be Covered

- VoIP vs other applications
- **Understanding and detecting typical network impairments in the enterprise network, and their impact on VoIP**
- VoIP Quality monitoring
- Data network techniques vs. VoIP requirements
- Wrap up

# Impairments

- Typical impairments that the network introduces in a VoIP environment.
  - Jitter
  - Packet Loss
  - Latency
  - Encoding/decoding (yea, this isn't really the "network")
- While not all impairments are equal, all of them have some negative impact on the user experience
- First though, we need to set some expectations about VoIP packets.

# VoIP Packets

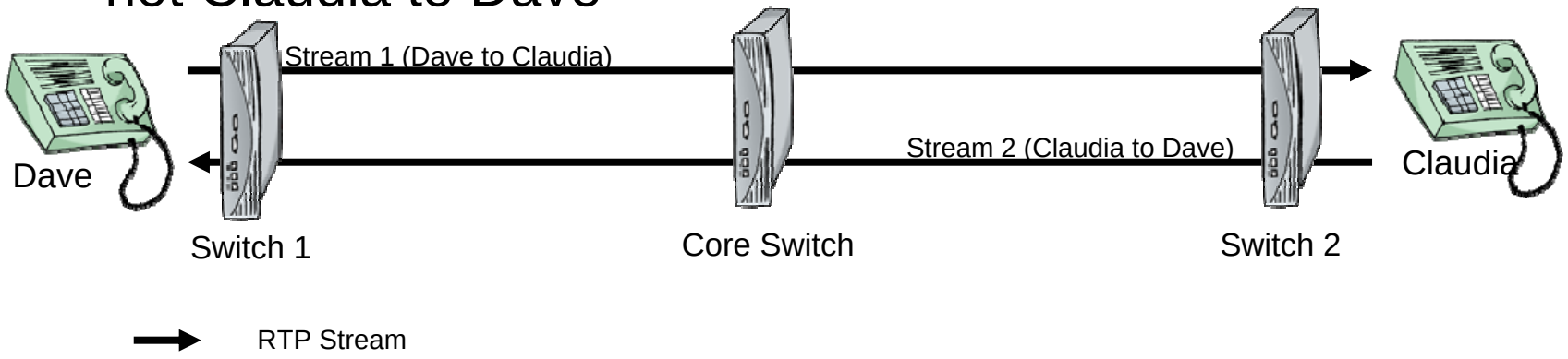
- When VoIP is sent across a network several things happen
  - Sender:
    - A Digital Signal Processor converts the analog voice sound into a digital set of bits and bytes using a coder-decoder (CODEC) algorithm
      - Each codec (and there are many) uses a different algorithm to do the conversion, we'll assume g.711
    - The codec creates small "Chunks" of audio 10ms long
    - Groups of 10ms chunks are packaged together in an IP packet and put on the network (2 groups of 10ms)
    - The phone then puts the packet on the network (one every 20ms)

# VoIP Packets (cont'd)

- Receiver:
  - Packets are expected every 20 ms, and are put in a queue (Jitter buffer)
  - When the jitter buffer has enough to work on (usually a few packets, it starts to work on them
  - Packets are taken from the queue they are pulled apart back into the right chunks (10ms for g.711)
  - The chunks are then "Decoded" by the codec algorithm to play back the audio stream to the earpiece
  - When the receiver is ready to play a sound to the earpiece and there is no packet to decode, there is a problem!
- This is the result of some impairment!
- Remember, VoIP is VERY deterministic on the network
  - You know you will get a packet of x length every y seconds, and when you don't either the conversation is done, or there is something wrong!

# More about VoIP

- A “conversation” is really two separate one-way streams of RTP packets
- Network performance is not usually the same in both directions, so impairments happen independently on each end
  - IE: I can have an impairment from Dave to Claudia, but not Claudia to Dave



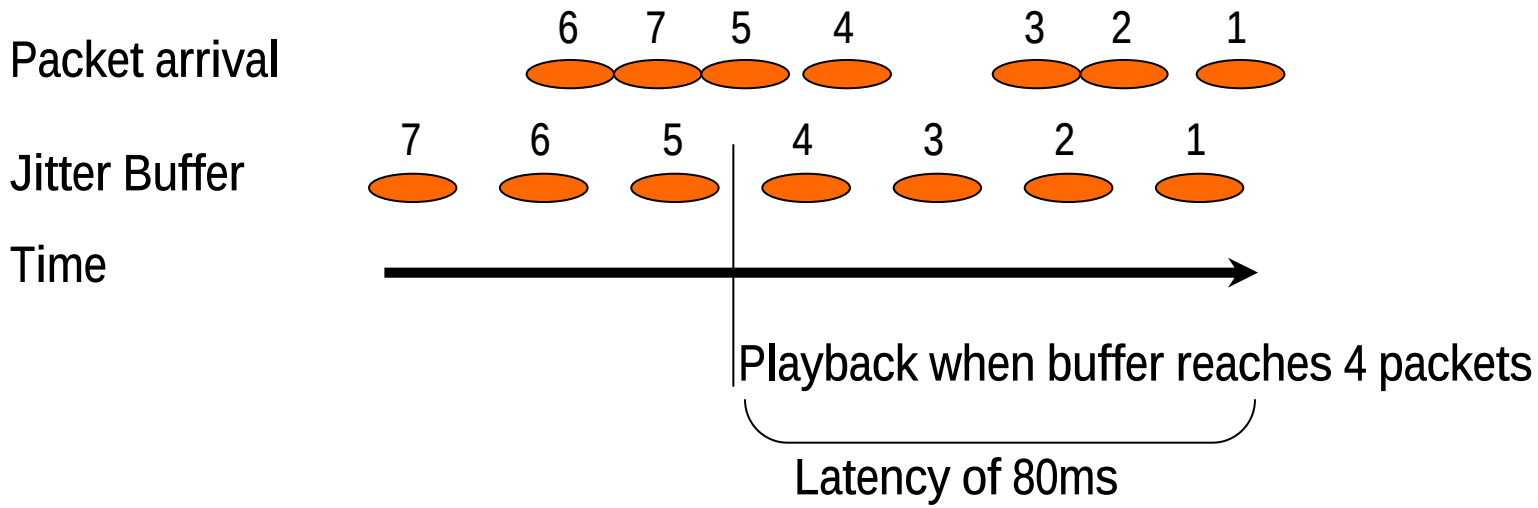
# Jitter

- VoIP Packets are expected at a specific interval
  - 20ms for g.711 in our example
- Jitter measurements indicate the “steadiness” of packet arrival at a point in the network
- Jitter means that packets are not arriving at expected intervals
  - Either they are arriving too quickly, or too late
- A jitter buffer is used to “smooth out” the jitter introduced by the network
  - Similar to the buffer of a streaming audio playback it saves up a few packets in a queue so that one is always ready when the codec needs it
- Jitter buffers introduce latency that is equal to the number of packets held in the queue prior to playback startup times the size of the packet in ms

# Jitter (cont'd)

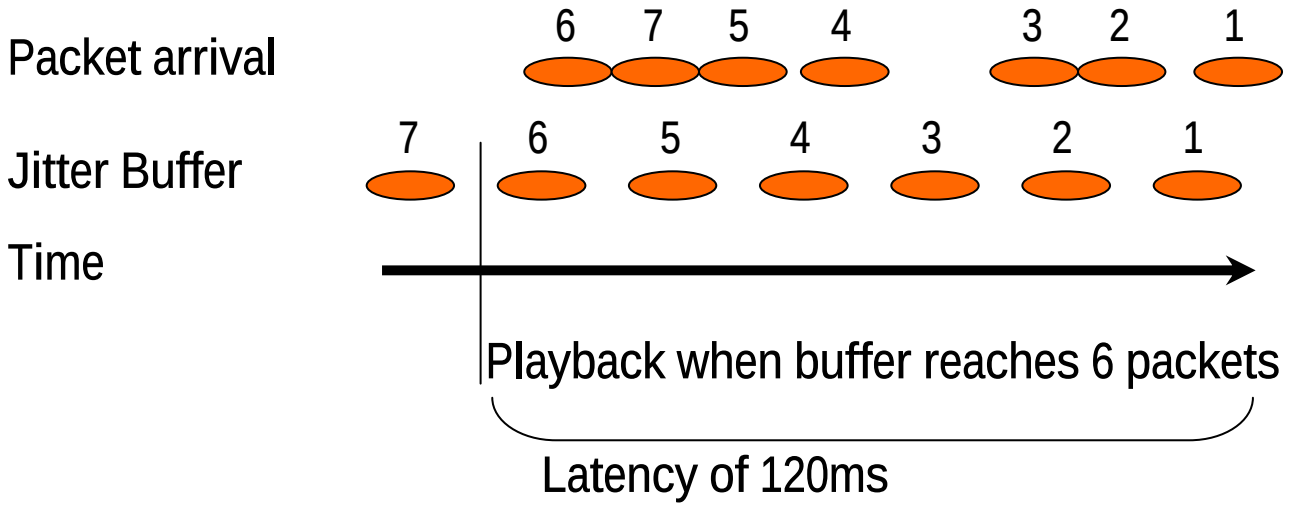
- When too much jitter occurs, there is no packet to "play" and silence happens in its place
  - There are some phones that use other mechanisms to avoid silence, but they only work for so many milliseconds before silence happens anyway
- More advanced jitter buffers adjust the queue length to optimize performance and minimize latency

# Jitter Buffer Functionality




- The jitter buffer will wait for a set amount of packets before playback, so that if a packet is lost, it has some “buffer” to wait for the next packet instead of playing silence
  - 4 packets in our example, causes 80ms of latency

# Jitter Buffer Functionality




- The jitter buffer will wait for a set amount of packets before playback, so that if a packet is lost, it has some “buffer” to wait for the next packet instead of playing silence
  - 4 packets in our example, causes 80ms of latency
- IF the jitter gets too high, it will “adapt” and store more packets before playback
  - 6 packets in our example, causes 120ms of latency

# Detecting Jitter

- Jitter (or the lack there of) can be easily measured by examining the arrival times of consecutive packets at any point in the network
  - Use the codec type in the RTP packet to determine when the next packet should occur (every 20 ms for g.711)
  - When packets arrive every 20ms, there is no jitter
  - When a packet arrives at 22 ms or 18ms, there is jitter
  - When a packet arrives at 22ms, then 22 ms later, then 20ms later, all 3 packets have jitter. Now you've lost 4ms forever.
- Before you jump to the conclusion that all is lost, you have to remember that the jitter buffer, by introducing some delay, may have been able to hide that jitter from the receiver

# Predominant Causes of Jitter

- Queuing delay
  - Large clumsy data packets getting in front of small speedy RTP packets
  - Implement, or check your QOS settings
- Route flapping
  - Re-routing due to a network failure
- Packet congestion
  - 10 pounds of packets in a 5 pound frame connection!
  - Again, implement QOS
  - Upgrade the circuit, or limit traffic
- Collisions
  - Not likely in today's switched network
  - But if you ARE using hubs, get rid of them

# Packet Loss

- The imperfect nature of an IP network means that sometimes packets get lost along the way
- TCP based applications (SMTP, SAP, HTTP) use retries to get ALL of the packets sent and received
- RTP, being UDP doesn't have that luxury
- RTP has added some functionality to it that ensures you know when you've lost packets
  - Sequence numbers, timestamps
- When you lose a g.711 packet, you've lost 20ms of "audio" to play back!
  - When you lose 5, you have a second to make up with something, hopefully not silence!

# Packet Loss Concealment

- Most good VoIP implementations have a Packet Loss Concealment (PLC) algorithm
  - Because of Jitter and loss, every so often there is no audio to play to the listener, and this is very discomforting
- PLCs “make up” a sound to play when there is nothing else to do
  - Replay the last packet
  - The CODEC may use information from previous and following packets to generate a seemingly appropriate sound
  - Sometimes it’s just “white noise”
- PLC can only do so much, typically effective for a few consecutive lost packets before the receiver is asking for something to be repeated!

# Detecting Packet Loss

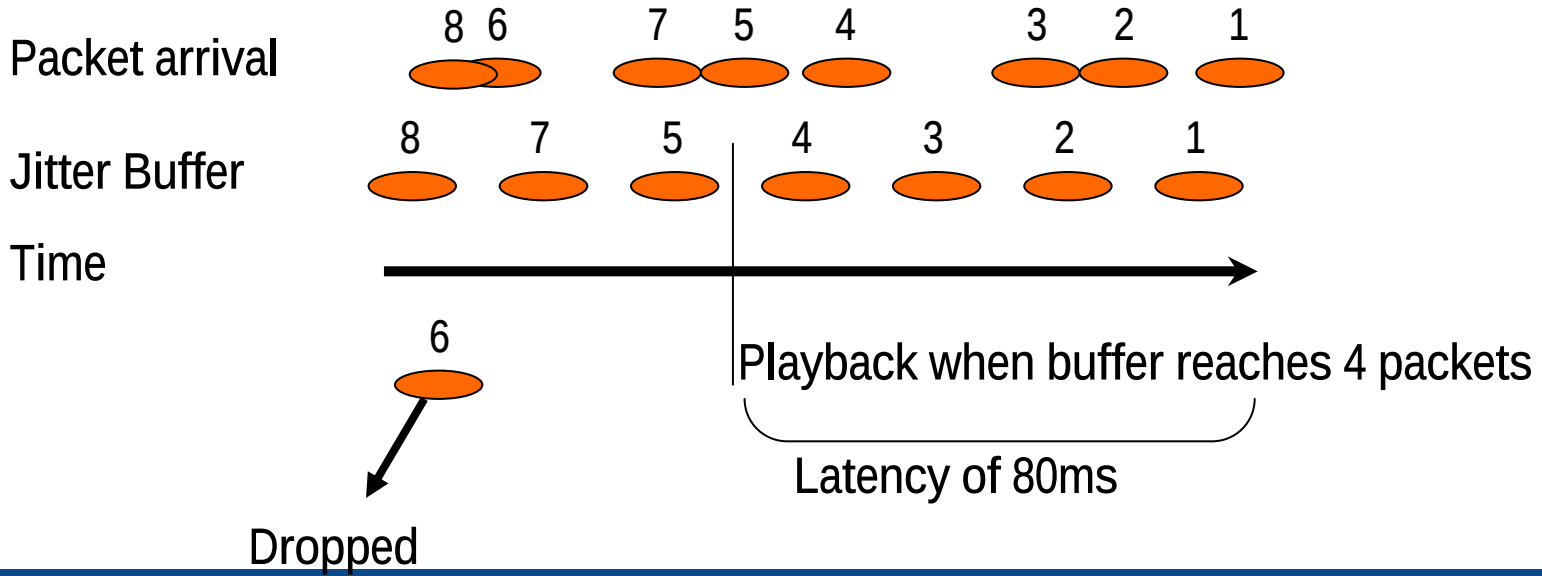
- The deterministic nature of VoIP enables monitoring tools to detect packet loss PER STREAM anywhere in the network
  - The sequence number increments with every packet
  - A combination of IP address, port number, SSRC number enable a tool (sniffer or otherwise) to separate each VoIP stream just from the header
  - Examine sequence numbers to ensure you aren't missing any!
  - Careful, there is a difference between losing packets and hanging up!

# Causes of Packet Loss

- Congestion
  - Frame Relay and other WAN networks will not always get all packets through
  - Ensure proper CIR, and monitor your Service Provider like a hawk
- Queuing discards
  - Routers will discard packets when queues fill up
  - Ensure that priority is set for VoIP, and queues can empty fast enough
- Buffer overflows
  - Independent of specific queues, the routers (Cisco) can run out of capacity to “hold” packets while it makes decisions about what to do with them
- Port speed mismatches

# One More Cause of Loss

- Packet loss due to Jitter Buffer Discards
  - The Jitter Buffer will “Toss” packets that arrive too late
  - This is almost certain to happen when a packet arrives both late and out of order
  - PLC or silence takes its place



# Latency

- Latency is a measurement of how long it takes to get a single packet to get where it needs to go
  - Typically measured in “round trip”
    - Ping measures round trip latency
  - VoIP, being two one-way streams, can’t use a round-trip measurement, it must be one-way latency
- Typically, one-way latency must be less than 150ms
  - Higher latency will be detected in conversation, and results in “talking over each other”
- The exception is VoIP used to carry Radio
  - Radio users are used to “abrupt” conversations with “push to talk” radios

# Measuring Latency

- Of the top 3 network performance problems for VoIP, latency is the most difficult to measure
- Round trip latency is easy to measure with a ping!
  - Even if it's not the right type of packet
- With RTP, Packet loss and Jitter can be measured anywhere in the path, or at the endpoints
- One-way latency requires the coordination of two points, and a strict time source (NTP)
- There are two methods
  - Sampled latency
  - Synthetic measurements

# Measuring Latency - Sampled

- Sampled latency takes a packet of actual VoIP traffic, and compares the time it arrived at one point in the network with the time it arrived at another point
  - Uses the IP address and port source/destination pair, SSRC# and sequence number as identification
  - Sniff packets at point A, and point B
  - Compare arrival times of two specific packets
  - Compare arrival times at a central management server
  - With multiple conversations, and a packet every 20ms (g.711) per conversation this is no easy task
  - Devices doing the sniffing must be NTP synced, AND careful that the arrival time is not negatively impacted by the time it takes to process the packet!
- Has the advantage of providing ACTUAL latency of REAL conversations, a great input to Mean Opinion Score (MOS)
- Challenge is knowing which two monitoring points will see the stream!

# Measuring Latency - Synthetic

- Synthetic latency generates fake traffic, and measures the latency (similar to a ping)
  - Endpoints must be established specifically for this purpose
    - Similar to Cisco SAA, or an appliance
  - Endpoints must be time sync'd (NTP)
  - Endpoint A sends a packet with a timestamp to endpoint B
  - Endpoint B does the measurement, and provides the result
  - Care must be taken to use an RTP packet, and set the right COS/QOS bits
- Gives a good estimate of latency of the real packets on that path
  - BUT think about how many paths you might have!

# Causes of Latency

- Everything that creates or handles a packet adds latency
  - The sending phone
    - DSP
    - CODEC delay
    - Packetization delay
    - Media Access delay
  - Switches
  - Routers
    - Queueing
    - Routing
  - WAN
    - Congestion

# Causes of Latency (cont'd)

- Everything that creates or handles a packet adds latency (cont'd)
  - Receiving phone
    - Jitter buffer
    - De-packetization
    - Codec
    - DSP

# The Latency Budget

- Phones (Sending/receiving) typically have a “Fixed” budget
  - They are almost always introducing the Same latency
- Switches will also introduce a small amount of fixed latency (media access on uplink)
- The laws of physics require that fiber, copper, and satellite connections introduce fixed amounts of latency
  - Sometimes this is significant!
- Routers introduce variable latency
  - Queuing, packet processing, buffering

IPCOMM2006
September 26-27 • Gaylord Opryland • Nashville, TN

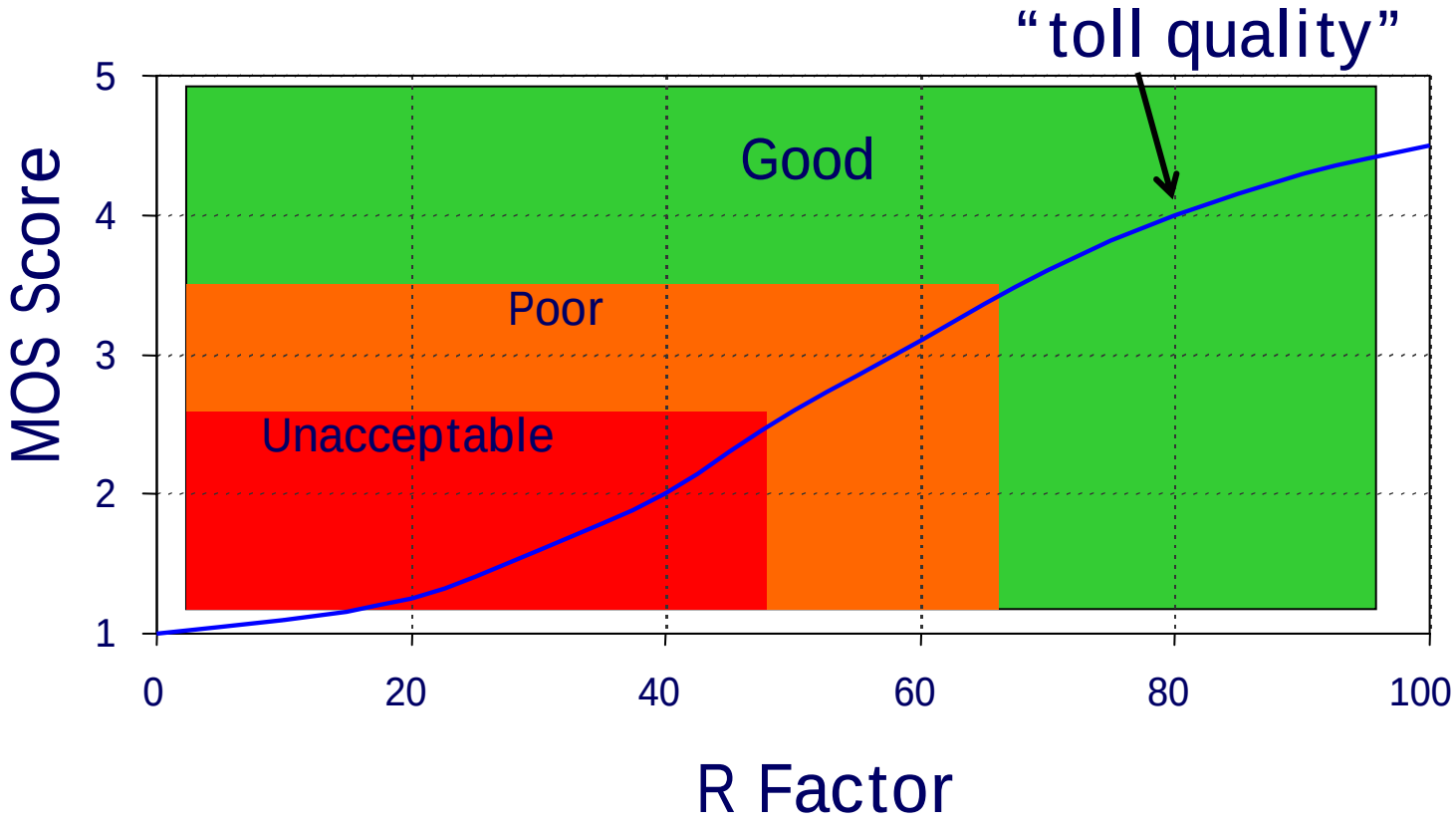
# What Will Be Covered

- VoIP vs other applications
- Understanding and detecting typical network impairments in the enterprise network, and their impact on VoIP
- **VoIP Quality monitoring**
- Data network techniques vs. VoIP requirements
- Wrap up

# The Relationship Between Quality and Performance

- On the PSTN the quality standard is a Mean Opinion Score (MOS)
  - Quality is a very perceptive measurement, one person may think it's fine, while another would be annoyed
  - Developed by the phone companies
  - 100 people "rate" the quality of a phone call on a 1-5 scale
  - Impairments added, and the call is re-rated
    - Over and over and over until a good baseline is established for each set of impairments
- MOS has been translated to the IP world through the use of an R factor and MOS score
  - Several standards exist for measuring the quality of VoIP and the impairments introduced by the network

# MOS Score Scale



# PESQ

- ITU standard (P.862) developed by Pystechnics
  - Injects a test sample into the network, and compares the resulting record to the original
- Pros
  - Can be used on VoIP or analog connections (or combination)
  - Very well coordinated with real perceptive MOS tests
- Cons
  - Time consuming, and computationally complex measurement
    - Cannot be done in real time, on real conversations
  - Injects more traffic on the network

# E-model

- The E-model (ITU standard G.107) provides a good mechanism to translate the performance metrics to a “quality” rating
- The “formula” applies packet loss, jitter, and latency to the behavior characteristics of a codec to calculate an R-Factor / MOS score
  - The measurements should be applied to EACH RTP stream independently
  - Note that one way latency must be used!
- Pros
  - Can be done in real-time on every conversation
  - Per minute results give very good granularity
  - The input measurements help network engineers direct troubleshooting
- Cons
  - Can only be used on IP networks, and show the degradation caused by IP network (Jitter, packet loss, latency!)

# MOS in a VoIP Network

- The effective "MOS" of a call in a VoIP network is primarily driven by:
  - Performance
    - Packet Loss
    - "Burstiness" of Packet Loss
    - Jitter Buffer Discards
    - Latency
  - Configuration
    - CODEC used
    - Packet Loss Concealment Used

# MOS in a VoIP Network

- The contentious nature of a IP network means that the performance metrics may change rapidly over the life of a call
  - Quality of Voice must be monitored at very granular level
  - Averages over the life of a call just won't do!
  - For example:
    - 2 min. of horrible quality at the end of a 40 minute call upsets the listener and they hang up
    - The average quality of the CALL is good (the 2 min of bad is averaged out by the previous 38 min. of good quality)
    - Operators who are looking through this data, don't see a bad call!
    - This is typical of "CDR based" MOS scores
  - Per minute MOS scores allow for avg, high, low analysis, and enables troubleshooters to find bad calls quickly and easily

IPCOMM2006
September 25-27 • Gaylord Opryland • Nashville, TN

# What Will Be Covered

- VoIP vs other applications
- Understanding and detecting typical network impairments in the enterprise network, and their impact on VoIP
- VoIP Quality monitoring
- **Data network techniques vs. VoIP requirements**
- Wrap up

# Traditional Data Network Management Techniques

- High degree of emphasis put on monitoring of components and sub-components
  - Routers, CPUs, Circuits, memory…
- Down stream event correlation
  - When a switch becomes unavailable, everything below it is unavailable
- Monitor application paths with synthetic or passive monitoring
  - Applications and services are centralized
    - Limited number of “application paths” across the network
- Agents, polling, and traps get the results needed to troubleshoot

# Traditional Data Network Management Techniques (cont'd)

- No need to actively monitor performance at the end-user device (PC)
- No need to monitor every transaction
  - The centralized nature of servers means that if one person is having a problem, all users have a problem (in a given location)
- Rarely relate statistics from a device to a single, specific end-user transaction to a larger problem

# Monitoring the Components

- VoIP relies heavily on multiple components to deliver on reliability, performance and quality
- The sum of the parts does not equal a good phone call
  - Each component may be within tolerances, but the cumulative effect of minor problems may impact the quality of a call
    - For example, if every component introduces the max latency without setting off an alarm, it may exceed the budget for latency
- VoIP performance/Quality must be measured holistically!
  - Measurements have to be taken at the call level

# Downstream Event Correlation

- The highly distributed nature of a VoIP infrastructure and the intelligence of a dial plan don't lend themselves to this
  - A PSTN interface being down doesn't mean that calls can't be made
    - Dial plans can re-route to other gateways or interfaces on that site or other sites
  - An IP PBX going down doesn't mean that phones can't make calls
    - Most IP phone systems have the ability for phones to register with another IP PBX
  - An IP WAN link doesn't mean that calls can't be made to the remote site
    - The dial plan may re-route calls over the PSTN

# Monitoring Decentralized Applications

- The application paths are highly decentralized
  - Phone to IP PBX
  - IP PBX to Phone
  - IP PBX to Gateway
  - Phone to Gateway
  - Gateway to IP PBX
  - Phone to Phone (and phones are everywhere!)
- There is no way to synthetically test each of these paths (much more than an $n^2$ problem)
  - Can only synthetically test a representative sample of the paths
  - Must use passive monitoring to monitoring the rest

# Use agents, polling, and traps

- Phones don't have agents with a wealth of performance information
  - Creates a problem of collecting the data from far more endpoints than any management system is capable of accomplishing
  - New standards will introduce quality metrics visible at the phone (RTCP-XR)
- Many IP-PBXs don't have processing power to run agents
- None will give accurate visibility into the end user experience
  - Phones don't have the spare cycles to do solid e-model quality metrics
  - PBXs are not in the call stream

# Monitoring End-user Devices and Transactions

- The peer-to-peer nature of VoIP conversations requires monitoring of each transaction individually
- Movement of phones means that you have to track each phone to be able to monitor its performance AND provide 911 services
  - "Hoteling", and other features make this more complicated
- The type of phone, it's firmware version, and its location can have a huge impact on the performance of calls to/from that phone
- Users tend to complain more about a specific phone call than with other transactions
  - It's also the case that if a transaction to a typical application server is slow, they will all be slow
  - With VoIP, one call may be great, the next one lousy

# Relating Performance of a Device to a Transaction

- The quality/performance of calls that traverse a gateway can be affected by both the PSTN and the IP network
  - The Gateway alone knows when the PSTN is giving it problems
  - The user is on the IP side
  - The Quality of the call from an IP perspective (MOS) must be correlated with the performance of the gateway interface to the PSTN to determine where any quality problem manifested itself
    - Echo, signal, latency, jitter… or some combination may be at fault

IPCOMM2006
September 25-27 • Gaylord Opryland • Nashville, TN

# What Will Be Covered

- VoIP vs other applications
- Understanding and detecting typical network impairments in the enterprise network, and their impact on VoIP
- VoIP Quality monitoring
- Data network techniques vs. VoIP requirements
- **Wrap up**

IPCOMM2006
September 26-27 • Gaylord Opryland • Nashville, TN

# Resources

- www.opticom.de/technology/pesq.html
- http://www.cisco.com/en/US/tech/tk652/tk698/technologies_tech_note09186a00800945df.shtml \
- http://www.nxtbook.com/fx/books/cisco/packet-3Q-05/
- http://www.qovia.com/resources/white_papers.htm

# Key Points to Take Home

- VoIP is different than any other application on the network
- VoIP is either viewed as strategic, or cost-saving
- VoIP is more important than any other application on the network
- The performance metrics that are most important to VoIP are Jitter, Packet Loss, and Latency
- Latency has a specific budget of <150ms
- Quality is subjective, but can be estimated with a MOS score
- Management tools must focus on the unique mechanisms of VoIP to be successful

IPCOMM2006
September 25-27 • Gaylord Opryland • Nashville, TN

# QUESTIONS?

**Contact:**
**Dave Chapman**
Dave@chapmanconsultants.com
**410 340 7597**

CHAPMAN CONSULTANTS