NTIA Special Publication SP-06-438

Proceedings of the International Symposium on Advanced Radio Technologies

March 7–9, 2006

Patricia J. Raush, General Chair
Kristen E. Novik, Publications

U.S. DEPARTMENT OF COMMERCE
Carlos M. Gutierrez, Secretary

John M. R. Kneuer, Acting Assistant Secretary
for Communications and Information

March 2006

DISCLAIMER

Certain commercial equipment, components, and software are identified to adequately present the underlying premises herein. In no case does such identification imply recommendation or endorsement by the National Telecommunications and Information Administration, nor does it imply that the equipment, components, or software identified is the best available for the particular applications or uses.

INTERNATIONAL SYMPOSIUM
ON
ADVANCED RADIO TECHNOLOGIES

March 7 - 9, 2006 · Boulder, Colorado

Sponsored by:

Institute for Telecommunication Sciences (ITS)
National Telecommunications and Information Administration (NTIA)

National Institute of Standards and Technology (NIST)

U.S. Department of Commerce, Boulder Laboratories

SAFECOM
National Department of Homeland Security (DHS)

RF Globalnet

University of Colorado Department of Interdisciplinary Telecommunications

2004 ISART Technical Committee:

Patricia J. Raush, General Chair, NTIA/ITS

Carolyn G. Ford, Technical Chair, NTIA/ITS

Fredrick Matos, NTIA/OSM

Stephen Berger, TEM Consulting

Tim Brown, University of Colorado

Robert Kubichek, University of Wyoming

Kate Remley, NIST

Robert Stafford, NTIA/ITS

CONTENTS

Page

The Role of the Regulator in Fostering Innovation

Christopher Haslett
Office of Communications (Ofcom)
Phone: +44 (0)20 7981 3750
Email: chris.haslett@ofcom.org.uk

Ofcom, the United Kingdom regulator, is required to secure the optimum use of the spectrum. Whilst the definition of "optimum" when applied to the spectrum could be debated, Ofcom has identified "encouraging innovation" as one of its major themes in pursuit of this requirement. The traditional "command and control" approach to assignment and licensing is not seen as an approach that encourages innovation. However, even prospective, and active, innovators are keen to see that levels of interference do not rise. Ofcom is exploring ways in which the spectrum usage rights granted to licensees can be made less prescriptive, particularly in terms of the technology deployed. The development of such rights should reduce the time for a new technology to be deployed. This involves two major elements: agreeing the nature of such rights to ensure that there is no ambiguity and developing appropriate methods of interference and coverage prediction to ensure that the bounds to these rights are not breached. This presentation gives details of developments made thus far in these areas.

Quality of Service Analysis of Site to Site IPSec VPNs for Real Time Multimedia Traffic

Jesús Pérez, Víctor Zárate, Ángel Montes, Carlos García
ITESM Campus Cuernavaca, Instituto de Investigaciones Eléctricas, Mexico.
{jesus.arturo.perez, vzarate, A00373556}@itesm.mx, cfgarcia@iie.org.mx

Abstract -- *This paper presents a quantitative analysis of the Quality of Service (QoS) offered by Virtual Private Networks (VPNs) based on the suite of protocols IPSec (IP Security) in a Computer Supported Collaborative Learning environment. Testing was done with routers and network nodes which held a communication channel through a videoconference. The objective is to analyze whether the IPSec VPNs are good enough to transmit real time multimedia traffic while protecting the information. Our results show that under stress traffic conditions the IPSec VPN could not offer an adequate QoS to the videoconference in terms of network latency, and that is why new alternative technologies need to be implemented in order to mitigate the degradation. The IPSec VPN technology is evaluated under the latency, jitter and packet loss parameters, which are the basic ones that determine the QoS in a point to point link.*

1. Introduction

Computer networks have evolved to the point where now it is possible to support multimedia applications (voice and video) over corporate networks and the Internet [1]. This has motivated the use of new technologies for the education field, as in the case of the Computer Supported Collaborative Learning (CSCL). The main characteristic of this architecture is the creation of virtual classrooms which allow students and teachers to interact in a real time collaborative way. All of this supported through IP infrastructure.

The use of videoconferences demands Quality of Service (QoS) requirements for voice and video traffic. QoS is defined as the network capacity to offer the best service for a selected traffic flow [2]. This means that voice and video require a minimum delay. Multimedia traffic transmission with a great jitter, low bandwidth and great loss could end up in an unacceptable communication [3].

Voice and video traffic is very sensitive to delay and the service offered by Internet Protocol called "best effort" is not adequate for this kind of traffic.

Nowadays is necessary to include QoS mechanisms over the link; but also protecting the information should be necessary either, especially when transmitting data over Internet or a shared WAN. This is required because Internet is a public network and is susceptible to many attacks. The IP protocol by itself does not protect the data over a public network: packets can be seen within the route towards the destination node, the IP address might be changed and also many other attacks can be mentioned.

In order to prevent and mitigate some attacks, the IPSec protocol suite was developed standard. Without any distinction, IPSec integrates security elements to the IP protocol such as: origin authentication, data integrity, confidentiality, no repudiation and anti-packet repetition [4].

In this context, our work's objective is to evaluate if a standard network infrastructure conformed primarily by routers (like a university campus or a small size network) can maintain the real time voice and video QoS parameters in a good degree while protecting them with an IPSec tunnel.

Many studies have been done to evaluate the IPSec performance but the shown results do not apply to our purposes since our network infrastructure includes routers which create the IPSec tunnels. Also, the data (in our scenario) to transmit is voice and video simultaneously over the same IPSec tunnel; the data is in real time and not buffered, generated by one videoconference using wired media. The results in [5] only focus in voice traffic and not in video traffic. In [6] the test scenario consists in a wireless network and the IPSec tunnel creation is based on desktop nodes. The same happens in [7] where no network layer equipment is included and also the test did not include voice or video traffic. The scenario in [8, 9] includes wireless equipment and no multimedia traffic considered for the results. In [10] the results include streaming voice and video, but this kind of traffic is not real time like the videoconference's traffic since the data is stored in a file before it is sent. In [11] the analysis is done with a different perspective because the evaluation is based on MIPS (Millions of Instructions Per Second) as the metric and not in terms of QoS parameters.

The paper is organized showing in section 2 basic concepts to understand the IPSec VPN operation. In section 3 we explain the QoS parameters. Section 4 shows the bench test where all the testing was implemented (videoconferences). In section 5 we explain our final results, while section 6 shows the conclusions achieved.

2. IPSec Virtual Private Networks (VPNs)

IPSec was designed and created by the IETF as the security architecture for the Internet Protocol IP. It defines the IP packet formats and infrastructure dedicated to provide authentication, data integrity, confidentiality and anti-packet repetition. IPSec might be implemented either in the origin/destination node or in the gateways/firewalls. In our case, the gateways were routers.

IPSec is based on two encapsulation protocols: ESP (Encapsulation Security Payload) and AH (Authentication Header). AH provides origin authentication, data integrity and anti-packet repetition. ESP also provides all characteristics mentioned above and additionally provides confidentiality through data encryption [12] [13].

ESP modifies the original IP packet inserting a new ESP header (after the IP header but before the data payload) and a packet trailer. The ESP header is not encrypted but a section of the trailer and the complete data payload are encrypted (figure 1). The packet authenticated part includes: ESP header, data payload and a trailer section.

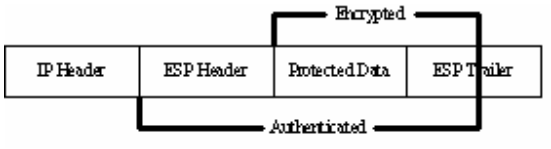

Figure 1. Encapsulated IP packet with ESP

When the destination node receives the IPSec packet, it receives it in clear: the Security Association (SA), the packet sequence number and the hash. This order is because of the same process of receiving that consists in three steps:

1. Sequence number verification.
2. Data integrity verification.
3. Decipher of information.

Before deciphering the information, which is a process that consumes lots of computational resources, the packets need to be checked to see if it did not delay (according with previous received packets) nor repeated. If the packet is valid, the next step is to verify the hash calculation and therefore check that the received information was not modified during transmission by a non authorized user or by a media failure. Finally the information is deciphered using the encryption shared key generated with IKE (Internet Key Exchange Protocol). At this moment the data is ready to be processed.

IKE is a hybrid protocol that gives different services to IPSec such as: IPSec peer authentication, security association agreement and key generation/regeneration for cipher algorithms used by IPSec. IKE negotiates the IPSec SAs. This process requires that IPSec peers get authenticated first with the help of digital certificates or pre-shared keys. After doing this, IKE can take further actions for the negotiation of IPSec SAs.

The next section explains the basic parameters needed to quantify the QoS performance.

3. Quality of Service (QoS)

There are some parameters which are necessary in order to quantify the performance. The usual ones are: latency, jitter and packets loss.

3.1. Latency

For our purposes we considered the latency as the time a packet takes to arrive to the destination network segment (one way latency). We started the count at the moment when packet is put by the origin node in its same network segment and finished the count when the packet arrived to the destination network segment. According to [14] the one way latency for multimedia traffic should not exceed 150ms.

3.2. Jitter

Or also known as packet delay variation, it is the average time that passes elapses between two consecutive packets at the destination node. The higher the jitter, the higher the quality degradation of voice and video will be. It is recommended that jitter should not exceed 50ms.

3.3. Packet Loss

It is the percentage of lost packets during transmission that were not received by the destination node. One of the main causes of packet loss is the network congestion over the links. Packet loss should not exceed 1% of the total transmitted packets.

When the above three parameter recommended limits are exceeded, it does not necessarily mean that the communication will be lost; it means that the quality for voice and video will be degraded in proportion to the exceeded recommended limit.

For this reason, in this paper we try to determine if a standard site to site IPSec VPN with different kinds of traffic is able to hold an appropriate QoS for multimedia traffic and in this way, to observe which are the degradation levels that the network experiments when data is protected with IPSec tunnels.

The next section explains the test scenarios implemented in order to evaluate the performance of the IPSec VPN; it explains the way the routers and the nodes were connected to establish the videoconference, the function of every node and the general aspects considered for the measurements.

4. Test Scenarios and Methodology

The test scenarios were designed according to a real university campus or a small/medium size company network. Our intention is not to simulate a network behavior: we wanted to create a real scenario to measure the performance seen in our topology.

Two scenarios were implemented in order to appreciate the behavior of the real time traffic under different traffic loads using an IPSec VPN: the first one consisted on four nodes; two for the videoconference and the remaining nodes generated TCP-FTP traffic through a large file download of 800Mbytes (figure 2). Videoconf. node A and Videoconf. node B held the videoconference. Node C connected to node E to download the file through a FTP session. We can say that in this configuration the network was an ideal scenario where no congestion occurred.

In test scenario #2 (figure 3) two more nodes were added to generate more traffic including file downloads using the HTTP protocol (with a Web page interface), we also injected high ICPM traffic load through sending large pings (10000 bytes size). In every network segment there was a FTP and HTTP server. Every client was connected with its server to create the connection. Both servers were used to manage the file download but with its correspondent protocol.

In both scenarios the IPSec VPN was implemented within two routers (A and C) creating one IPSec tunnel to protect exclusively the voice and video packets in both ways (figure 2 and 3). All FTP, HTTP and ICPM traffic was not protected with the VPN. Every WAN link was set up to the speed of 1Mbps by using the router's serial interfaces. Two nodes (Videoconf. node A and Videoconf. node B) held a videoconference with VIGO VCON proprietary consoles, cameras and microphones. The protocols for multimedia traffic were G.722 (for voice) and H.263 (for video).

The IPSec cipher specification was AES as the encryption algorithm, HMAC-SHA as the integrity mechanism, ESP encapsulation in tunnel mode, IKE as the key interchange protocol and the IPSec router authentication was made with pre-shared keys.

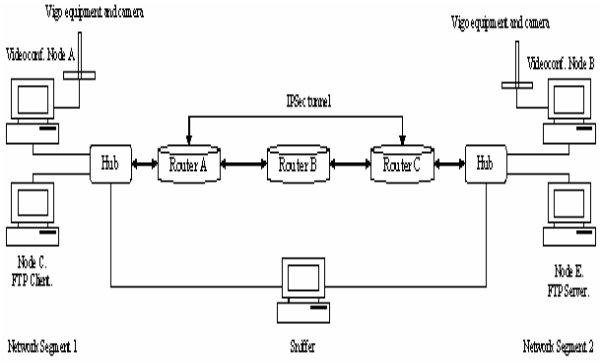

Figure 2. Test scenario #1

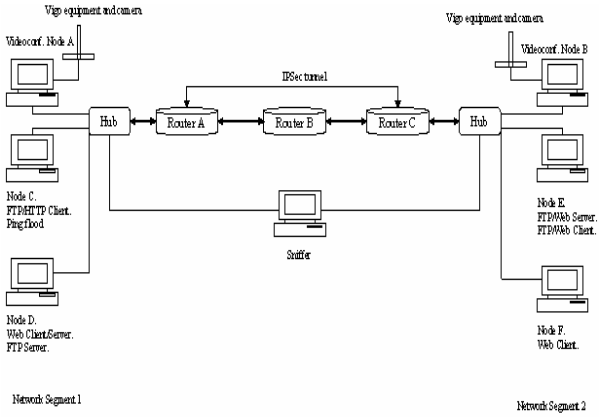

Figure 3. Test scenario #2

Testing consisted on capturing the UDP voice and video packets that traveled from Videoconf. node B to Videoconf. node A with the use of the Etherpeek NX sniffer. We only considered for the results the UDP packets coming from the videoconference. The remaining traffic (FTP, HTTP and ICPM) was discarded since it was just injected for increasing the WAN traffic. The sniffer node captured the incoming and outgoing traffic with two network adapters (NICs).

In order to perform the QoS evaluation, we established twenty real time videoconferences. Every videoconference lasted 2 minutes, during that time all multimedia packets were captured making a total of 40 minutes of videoconference packets captured.

The first ten videoconferences were implemented under test scenario #1, the other ten videoconferences under test scenario #2. In every test scenario, five videoconferences were implemented without the IPSec VPN and five using the IPSec VPN. The results obtained were averaged to reduce the error margin and are shown next.

5. Results Analysis

5.1. Latency

The results of the packets latency are shown on table 1.

Table 1. Results of packets latency

		Latency	
		Voice (ms)	Video (ms)
Scenario 1	With VPN	68.7	71.29
	Without VPN	66.82	63.69
Scenario 2	With VPN	252.18	264.77
	Without VPN	81.26	95.47

In test scenario #1 the latency never exceeded 150ms with and without IPSec VPN. Voice and video packets have close (but different) results producing a good videoconference quality. The main reason of this difference is the encryption process. Encryption demands more CPU and memory space in order to produce protected packets. The AES algorithm did not affected in a considerable way the packet latency but monitoring the router's CPU it used 30% of it. This percentage is vast in terms of resources usage since the router main task is not exclusively encrypting packets, the router also has to route packets towards its destination, create and maintain routing tables and many others.

In test scenario #2 the difference is vast since the traffic load conditions were different. Without the IPSec VPN, the voice and video packets reached around 81 and 95ms of latency respectively. But having the VPN implemented the latency went to 252 and 264ms approximately for voice and video. We attribute this behavior to the traffic load since the multimedia packets have to compete for the access to the serial link. Having a greater traffic load to serve, the packets need to wait more time (buffered) to get access to the medium. The multimedia packets, although they have different needs (latency, jitter and loss), are treated in the same way as a HTTP, TCP or as ICMP packets. There was no preference for this type of traffic.

The encryption process also could be affected since the router's CPU has to process more traffic. Meanwhile the CPU has more tasks to implement and the CPU throughput would be lower, making the encryption to take more time. The encryption process is relevant for increasing the packet latency but it is not the main reason for this latency behavior in test scenario #2.

In general, the video packets took more time to be sent; the reason is because video packets are longer than voice packets. A voice packet occupied 538 bytes, but a video packet average size was 1300 bytes. The serialization delay which is the time that the router takes to put the bits in the medium tends to increase while the packet size gets longer.

5.2. Jitter

The results of the packets jitter are shown on table 2.

Table 2. Results of packets jitter

		Jitter	
		Voice (ms)	Video (ms)
Scenario 1	With VPN	59.96	28.72
	Without VPN	59.97	32.84
Scenario 2	With VPN	60.07	30.28
	Without VPN	60.12	30.98

The jitter for voice packets with and without VPN was higher (almost 10ms) than the recommended 50ms. For the video packets, the jitter with and without VPN remained around 30ms. It can be seen that the encryption process did not influence in the jitter parameter for G.722 and H.263 traffic.

Despite of the 60ms voice jitter average, the end user did not notice a bad voice quality for two main reasons: first, 60ms are not too far away from 50ms and second, the end user listened the voice with hardware dedicated consoles. These consoles have memory for buffering that could compensate the changes in the arrival time. Having buffers for jitter control, the asynchronous arrival time became synchronous, therefore improving the voice and video quality. We can attribute the results similarities in both scenarios due to the constant traffic injection rate

and because the FTP and HTTP sessions did not have traffic peaks.

5.3. Packet Loss

The results of the packet loss are shown on table 3.

Table 3. Results of packet loss

		Packet Loss	
		Voice (%)	Video (%)
Scenario 1	With VPN	0	0
	Without VPN	0	0
Scenario 2	With VPN	0	0.01
	Without VPN	0	0

The third important parameter considered was the packet loss. As seen in figure 6 the packets loss was almost null in both test scenarios. The percentage was obtained based on the total amount of packets transmitted by the origin node towards the destination (videoconference nodes).

We can attribute the low increment of packet loss to the IPSec sliding window that detects repeated and out of time packets (similar to the TCP sliding window). All incoming packet must fit inside the sliding window. The traffic load and the encryption process increased the packet latency, therefore there were probabilities that the incoming packets would had not fit in the IPSec sliding window at the correct time. If so, the packet was discarded. Another important factor of the low packet loss was the router's serial interfaces matched speed. There was not speed mismatch between the three routers.

6. Conclusions

Based on the obtained results we can conclude that the QoS in a videoconference using IP infrastructure is affected in two parameters: latency and loss when using IPSec tunnels. The main two reasons of this behavior are the encryption process and the traffic load. Encryption requires great amount of CPU and memory, for this reason the router's manufacturers recommend the installation of VPN accelerator cards which are dedicated exclusively to data encryption/decryption, and in this way, the router is not involved in any other tasks other than routing information, thus increasing the router performance.

With a great amount of traffic, the router buffers the multimedia traffic until the serial link is no longer used. For this reason the latency increments depending on the traffic load. In order to decrease the latency, preferential treatment must be given to this kind of traffic over the remaining traffic.

The jitter parameter was not affected by the VPN. Even though the metric remained a little over the ideal limit with and without VPN, it did not affect the videoconference quality in a visible or audible way. The packet loss percentage changed not much in our test scenarios having or not having the VPN implemented since there was not any interface speed mismatch.

From above reasons, we could deduce that is feasible to implement IPSec VPNs for our collaborative learning architecture over a standard campus network (where the link's speed remain around 1 Mbps) but with medium traffic loads. Additional considerations should be taken when applying to bigger networks such as an Internet service provider, since the behavior is different, having peak times and a higher amount of traffic, resulting in a non-optimum solution. If the videoconference was set in peak times of flow without QoS mechanisms, the quality could be affected seriously.

In highly saturated networks it is necessary to use techniques able to protect and prioritize the information in order to make the traffic transmission secure without affecting the QoS parameters.

7. References

[1] Minoli Daniel, Minoli Emma, "Delivering Voice over IP Networks", Wiley Computer Publishing, 1st edition. 1998. ISBN 0-471-25482-7.

[2] Cisco Systems, "Wireless Quality-of-Service Deployment Guide", 2000, www.cisco.com

[3] Monsour Yishay, Patt-Shamir Boaz, "Jitter control in QoS networks", *IEEE/ACM Transactions on Networking (TON)*, Vol. 9, Issue 4, August 2001.

[4] Naganand Doraswamy, Harkins Dan, "IPSec The New Security Standar for the Internet, Intranets, and Virtual Private Networks", Prentice Hall PTR Internet Infraestructure Series, 1999, ISBN 013011898.

[5] Barbieri R, Bruschi D, Rosti E, "Voice over IPsec: analysis and solutions", *Proceedings of the IEEE Computer Security Applications Conference*, 9-13 Dec. 2002, pp.261 – 270

[6] Wei Qu, Srinivas S., "IPSec-based secure wireless virtual private network"; *IEEE MILCOM 2002 Proceedings* , Vol. 2 , 7-10 Oct. 2002, pp.1107 – 1112.

[7] Khanvilkar S., Khokhar A., "Virtual private networks: an overview with performance evaluation", *IEEE Communications Magazine*, Vol. 42 , Issue: 10, Oct. 2004, pp:146 – 154.

[8] Khatavkar D., Hixon E.R., Pendse, R.,"Quantizing the throughput reduction of IPSec with mobile IP", *Circuits and Systems, MWSCAS-2002*, Vol. 3, 4-7Aug. 2002, pp:III-505 - III-508.

[9] Hadjichristofi G.C., Davis N.J. IV, Midkiff S.F., "IPSec overhead in wireline and wireless networks for Web and email applications", *Conference Proceedings of the IEEE International Performance, Computing, and Communications*, 9-11April 2003, pp. 543 – 547.

[10] Al-Khayatt S., Shaikh S.A., Akhgar B., Siddiqi J., "Performance of multimedia applications with IPSec tunneling", *IEEE Proceedings of International Conference on Coding and Computing*, 8-10 April 2002, pp.134 – 138.

[11] Elkeelany O., Matalgah M.M., Sheikh K.P., Thaker M., Chaudhry G., Medhi D., Qaddour J., "Performance analysis of IPSec protocol: encryption and authentication", *IEEE International Conference on Communications*, ICC, Vol.2, 28April-2May 2002, pp.1164 – 1168.

[12] Kent S. "IP Authentication Header", RFC 2402, IETF Network Working Group, 1998. http://rfc.net/rfc2402.html

[13] Kent S. "Encapsulating Security Payload", RFC 2406, IETF Network Working Group, 1998.http://rfc.net/rfc2402.html

[14] Szigeti Tim, Hattingh Christina, "End-to-End QoS Network Design: Quality of Service in LANs, WANs, and VPNs". Cisco Press, 1st edition, 2004, ISBN 1-58705-176-1.

Introduction to Objective Multimedia Quality Assessment Models

Carolyn Ford, Arthur Webster
Institute for Telecommunication Sciences
National Telecommunications and Information Administration
U.S. Department of Commerce
Boulder, Colorado USA
Phone: (303) 497-3728, fax: (303) 497-6982
cford@its.bldrdoc.gov, awebster@its.bldrdoc.gov

The transmission of multimedia signals over wireless channels has increased exponentially in the past decade. The widespread use of digital technology for the transmission of audio and video signals has led to the need for objective quality assessment methods based on human perception. In particular, the distribution of multimedia signals over wireless links to devices such as laptops, PDAs, and cell phones is widespread. Manufacturers can use objective models to improve products and analyze deployment. Service providers can use objective models to monitor the quality of service they provide. This paper is an introduction to the concepts of multimedia quality assessment models and the design of subjective tests for objective model validation.

1. Introduction

The transmission of multimedia signals over wireless channels has increased exponentially in the past decade. As higher bandwidths become available, applications to fill the bandwidth have increased proportionately. What used to be the realm of analog broadcast television is rapidly becoming dominated by digital transmissions. Digital transmission is used for broadcast applications (e.g., television) as well as point-to-point transmission (e.g., cell phone) of audio and video content.

The widespread use of digital technology for the transmission of audio and video signals has led to the need for objective quality assessment methods based on human perception. Manufacturers can use objective models to improve products and analyze deployment. Service providers can use objective models to monitor the quality of service they provide. Buyers can use objective models to specify requirements, make purchasing decisions and analyze the quality of service received.

Analog methods, such as signal to noise ratio (SNR), while still important, are not sufficient to measure the quality of delivered audio and video signals. Standards for subjective assessment and objective methods for assessing analog signals have been in use for decades, but new subjective and objective methods are needed for both accuracy and ease of use [21]. This paper is an introduction to the concepts of multimedia quality assessment models and the design of subjective tests for model validation.

1.1 Background

Since the early 1990s, the International Telecommunication Union – Telecommunication Standardization Sector (ITU-T) Study Groups 9 and 12, the International Telecommunication Union – Radiocommunication Sector (ITU-R) Study Group 6, and the Alliance for Telecommunications Industry Solutions (ATIS) Performance, Reliability, and Quality of Service Committee (PRQC), and their predecessor organizations[1] have been developing standards for objective measurement of the perceived quality of digital video, speech, and audio. In recent years, perceptual models for the objective measurement of audio quality and video quality have been standardized and put into widespread use. ANSI T1.518-2003 [2] and ITU-T Recommendation P.862 [15] address objective audio quality measurement for narrowband voice and ITU-R Recommendation BS.1387-1 [3] addresses wideband audio. ANSI T1.801.03-2003 [1], ITU-T J.144 [13], and ITU-R BT.1683 [5] provide objective measurement methods of digital video quality for television applications. J.144 and BT.1683 each recommend the same four objective models: the NTIA general model (USA), the British Telecom model (UK), the Yonsei University model (Korea), and the CPqD model (Brazil). These video quality models were validated by the Video Quality Experts Group (VQEG) [www.vqeg.org][2]. Some of these standardized methods are available in commercial products. The NTIA video quality model is available in software with a free evaluation license[3]. It is also described in an NTIA

[1] ITU-R Study Group 6 was formed from ITU-R Study Groups 10 and 11. PRQC was formerly known as T1A1.

[2] The VQEG is an unofficial group formed from video experts of ITU-T Study Groups 9 and 12, ITU-R Study Group 6, industry, and academia.

[3] http://www.its.bldrdoc.gov/n3/video/vqmsoftware.htm

Report [22]. Contact information for the other video models can be found in J.144.

Until recently the development of an objective measure of overall multimedia quality has not been addressed. Multimedia is defined here as the combination of audio and video in the communication of information. Some work has been done combining audio and video quality primarily in subjective experiments [6][7][8][9][10]. The "mapping from the one-way audio and one-way video quality, as derived from audio only and video only subjective experiments, to the one-way overall audiovisual quality [was studied]. Four different laboratories found similar mapping results despite the fact that experimental conditions were quite different." [17]. The laboratories that did this work in the 1990s were Bellcore (USA), KPN Research (Netherlands), NTIA/ITS (USA), and France Telecom/CNET (France). These laboratories did subjective experiments to explore the relationships between audio and video quality. New tests are being conducted to better address audiovisual quality assessment in current multimedia applications.

In 2003 ITU-T Study Group 9 approved Recommendation J.148 entitled "Requirements for an objective perceptual multimedia quality model" [14]. VQEG and the Joint Rapporteur Group on Multimedia Quality Assessment (JRG-MMQA)[4] are currently working on a validation test covering video quality assessment models for multimedia applications. A test plan [19] for VQEG's first phase of multimedia testing (video only) has been approved and testing should begin in 2006. NTIA/ITS is participating in the VQEG tests and is, in addition, conducting independent tests to support our model development.

2. Definitions

Clip - Digital representation of a scene (defined below) that is stored on computer media.

Codec - Abbreviation for a coder/decoder or compressor/decompressor.

Common Intermediate Format (CIF) - A video sampling structure used for video teleconferencing where the luminance channel is sampled at 352 pixels by 288 lines [11].

Feature - A quantity of information associated with, or extracted from, a spatial-temporal sub-region of a video stream (either an original video stream or a processed video stream).

Frame – One complete television picture.

Hypothetical Reference Circuit (HRC) - A video system under test such as a codec and/or digital video transmission system.

Input Video - Video before being processed or distorted by an HRC (see Figure 1). Input video may also be referred to as Original Video.

Institute for Telecommunication Sciences (ITS) - The research and engineering laboratory of the National Telecommunications and Information Administration, U.S. Department of Commerce.

Mean Opinion Score (MOS) - The average subjective quality judgment assigned by a panel of viewers to a processed video clip.

Multimedia Quality (MMQ) - acronym

Multimedia Quality Metric (MMQM) - An overall measure of video impairment reported by a particular multimedia quality model, either for an individual MM clip (Clip MMQM), or for an HRC (HRC MMQM).

National Television Systems Committee (NTSC) - The 525-line analog color video composite system adopted by the US and other countries (excluding Europe) [18].

Original Video - Video before being processed or distorted by an HRC (see Figure 1). Original video may also be referred to as input video since this is the video input to the digital video transmission system.

Parameter - A measure of video distortion that is the result of comparing two parallel streams of features, one stream from the original video and the corresponding stream from the processed video.

Processed Video - Video that has been processed or distorted by an HRC (see Figure 1). Processed video may also be referred to as output video since this is the video output from the digital video transmission system.

Quarter Common Intermediate Format (QCIF) - A video sampling structure used for video teleconferencing where the luminance channel is sampled at 176 pixels by 144 lines[11].

Scene - A sequence of video frames, with or without accompanying audio.

Video Graphics Array (VGA) – A video format with pixel resolution of 640 x 480.

Video Quality Metric (VQM) - An overall measure of video impairment reported by a particular VQM model, either for an individual video clip (Clip VQM), or for an HRC (HRC VQM). VQM is reported as a single number and usually has a nominal output range from

[4] The JRG-MMQA is an official body of the ITU and is formed from members of ITU-T Study Groups 9 and 12.

zero to one, where zero is no perceived impairment and one is maximum perceived impairment.

3. Multimedia Quality Assessment

Multimedia quality measurements (i.e. audio and video measurements) characterize either the absolute quality or the degradation in quality of a multimedia signal that has been processed through a system. The system consists of an encoder/decoder pair and a digital channel that connects them. This end-to-end path for a multimedia signal is called an HRC (see definitions). To determine the effect of the system on the multimedia quality, the signal is sampled and calculations are made that have been determined to be significant to perceived multimedia quality. One configuration for making these measurements can be seen in Figure 1.

If the measurement system has access to the full original signal and the full processed signal, the measurement system is referred to as Full Reference (FR). If the measurement system utilizes only an extraction of features from the original signal shared over an ancillary data channel, the system is referred to as Reduced Reference (RR). Figure 1 shows an RR system. Methods also exist which make estimations of quality based upon only the processed signal. This class of measurement method is referred to as No Reference (NR). This paper will describe an example of the RR method. See [12] for a complete description of FR, RR, and NR methodologies.

In the RR measurement method, features are extracted from both the original and processed signals. The features are used to calculate quality parameters. Some of the parameters that are used to produce a quality measurement include the comparison of extracted features to characterize some aspect of the signal (e.g., edge magnitude, edge angle, motion, and local color for video). Features are compared to produce parameters that quantize certain aspects of quality (e.g., blurring, blocking, unnatural motion, color changes).

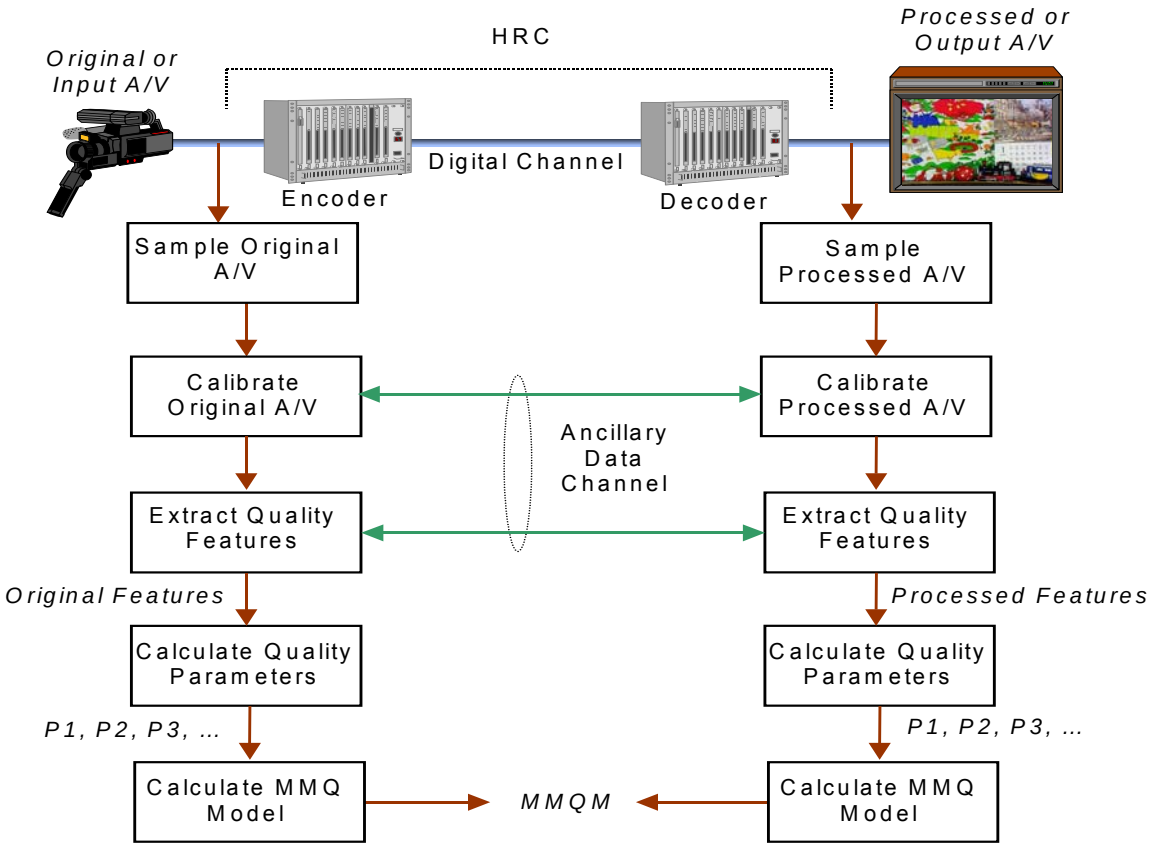

Figure 1. Steps required to compute multimedia quality metric.

3.1 Multimedia Models

The role of an MMQ Model is to predict a human subject's reaction to degradation in quality of a multimedia signal that has been processed through an HRC. Using the measurement process in Figure 1, the quality parameters are calculated. The next step for an MMQ Model is to make a determination of what a person's perceived quality of the processed video

would be. This determination is based on the quality parameters and the comparison of parameters calculated from the source sequence with those calculated from the processed sequence.

Figure 1 depicts a multimedia signal being processed as a single entity; however, the channel may affect the audio and video signals differently, so it is necessary to process each signal (audio and video) separately using the process shown in Figure 1. After the audio and video have been processed separately, their parameters must be combined appropriately. This is the challenge of the multimedia quality effort.

3.1.2 Requirements for Multimedia Models [14]

Figure 2 depicts the basic form of a multimedia model. The separate audio and video models provide inputs to the multimedia model. The focus of our research is to define the form of the multimedia quality integration function. The integration function applies specific rules to the information provided by the audio and video inputs. The form of these rules will be based on data derived from subjective quality experiments. The aim is to produce a set of integration rules that enable the multimedia model to accurately predict human quality perception of systems and services under test. Therefore the validity of the model must be shown by comparing the performance of the model against quality ratings obtained from subjective tests for a range of test materials. Subjective testing is discussed in section 4.

The multimedia quality integration function contains three primary input modules. Two modules provide predictions of audio quality and video quality for some multimedia service, and a third provides an indication of the differential delay between the audio and video sources. Standardized models will be used as the quality inputs to the multimedia model. There is ongoing research in developing perceptual quality models for audio and video with full reference, reduced reference, and no reference capabilities. A fourth input allows the model to accommodate any task-dependent influences that may impact quality perceptions. One role of the task input module is to represent the degree of interactivity associated with the multimedia service. This may also be a more comprehensive control, such as choosing specialized audio and video quality models that are best suited to the current task (e.g., a speech quality model suits a videoconferencing task). The multimedia model will be based on generic rules that capture human perceptions of audiovisual quality.

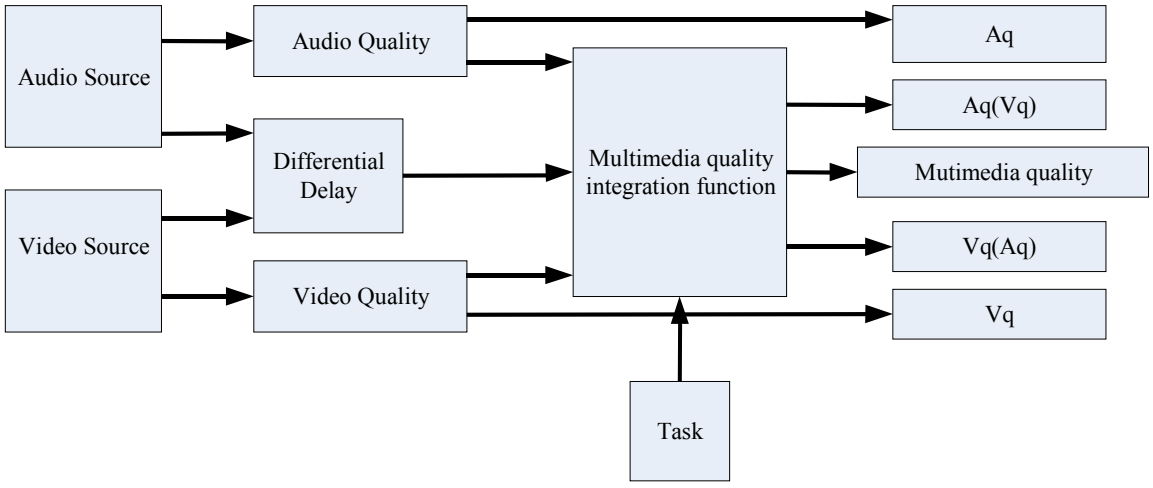

Figure 2. Basic components of a multimedia model

3.1.3 Multimedia Quality Integration Function

The multimedia quality integration function needs to accommodate human perceptual and cognitive processes active in the formation of quality judgments of audiovisual services. Data obtained from subjective quality tests will be used to define the integration function. The role of the integration function is to accept inputs from the audio and video models, and apply some predefined rules to the incoming data to produce a multimedia quality prediction.

The integration function will account for human perceptual responses to multimedia services. In particular, the integration function will apply rules that

represent elementary perceptual processes present in subjective quality assessment of multimedia. The complete multimedia model provides five outputs. The primary output is a predicted measure of overall multimedia quality. Four subsidiary outputs provide predictions of perceived quality for the audio (denoted Aq), video (denoted Vq), audio accounting for any influence the video may have (denoted Aq(Vq)), and video perceptual quality accounting for any influence the audio may have (denoted Vq(Aq)).

4 Subjective Testing

As discussed in section 3, objective quality assessments are based on data derived from subjective quality experiments. The validity of any model must be shown by comparing the performance of the model against quality ratings obtained from subjective tests for a range of test materials. This section presents some background on subjecting testing, and describes the design and execution of a subjective test for video for multimedia applications, as an example.

Subjective testing employs human subjects to rate multimedia quality. The results are used to train and test the objective measurements calculated by the models in order to improve audio/video codec quality, to analyze the interaction between network loading and received quality, and to compare the performance of different displays, coders, decoders, and networks. When carefully performed, subjective testing provides "truth data" – people's actual opinions. Care must be taken, however, or the subjective testing will result in misleading or incorrect conclusions. ITU-R Rec. BT.500 [4] and ITU-T Rec. P.910 [16] address the methodology for subjective testing of video, and the ITU-T has published several of VQEG's test plans and reports in a 200-page tutorial [20]. The VQEG tutorial addresses all aspects of designing and executing a multimedia subjective test, including test material, test methods, viewing conditions and data analysis. Researchers planning subjective testing are encouraged to review these documents. The ITU Recommendations describe several different options for subjective testing in order to answer different kinds of questions.

4.1 An Example Video Subjective Test

ITS is performing a subjective video test as the first step in the multimedia test process. The test described is the first of a series of multimedia subjective tests that will explore the relationships between the quality parameters for audio and video, as discussed in section 3. The data collected from this subjective test will be used initially to evaluate the effect of video resolution on perceived quality. The results will also contribute to the Vq function parameter in the ongoing multimedia quality assessment modeling effort and multimedia quality integration function research. The subsequent subjective tests will combine video and audio.

This test employs the Absolute Category Rating (ACR) method as documented in P.910 and as amended in the VQEG's multimedia testplan [19]. The subject rates both processed clips and their associated original clips; however, this method requires the test subject to rate clips individually, with no direct comparison to the original clip. Original and processed clips are combined into a set, and the clips in the set are shown in a randomized order.

4.1.1 Video Clip Selection

ITS collected a large set of high quality original scenes. The original video clips were transmitted through different configurations of video transmission systems to create processed video clips. The resulting set of video clips, original and processed, spanned a broad range of quality from excellent to bad. Each scene was scaled to the resolutions VGA, CIF and QCIF. VGA resolution is approximately that of a television show viewed at full resolution on a PC monitor. CIF resolution is approximately one-fourth the size of VGA, and is typically used by PDAs and some web sites. QCIF resolution is one-fourth the size of CIF and is typically used by cell phones and video on web sites intended for very low bandwidth connections. Each subject was shown the same scene content in each resolution, in varying orders of resolution (VGA-CIF-QCIF, CIF-VGA-QCIF, etc.). Subjects viewed each video clip in turn and indicated their opinion of the video quality.

The processed clips were originally intended to be viewed on a television monitor in NTSC format. These sequences had previously been rated in subjective tests to generate mean opinion scores (MOS). These subjective scores enabled ITS to select original video scenes that spanned a wide range of coding difficulty, from nearly still scenes containing very little detail, to complicated scenes containing multiple scene cuts and fast motion. The video clips were converted to VGA, CIF and QCIF using the approximate area a viewer would see on a television screen. The clips were chosen so that the viewed set would contain clips with a uniform distribution of MOS scores.

Original video sequences should be selected carefully. The content displayed should be visually different, so that subjects do not become bored. Content should cover a wide variety of material spanning a wide range of difficulty. For example, if all of the source scenes are head and shoulder shots with very little motion, then the codecs will respond identically to all content, and little is learned from including multiple scenes. The better approach is to

examine the amount of motion and detail in the source scenes, and then choose scenes that contain different amounts of motion and detail. For example, a head and shoulders videoconferencing scene might contain very little spatial detail and be nearly still, while a movie trailer will contain rapid motion, multiple scene cuts, and highly detailed scenery.

Processed video sequences and systems to be tested should be selected carefully to cover a wide range and type of quality degradation. Taken together, all of the original and processed video clips should evenly span a moderate to wide range of quality

4.1.2 Viewing Environment

The test is administered using personal computers (PC's). Video clips are displayed at random on the PC monitor, and the subjects are given unlimited time to rate each clip before the next one is shown.

Any number of identical PC/monitor combinations can be used, since the test is administered individually, and at the test subject's pace. We use two PC/monitor setups. The PC system must be capable of playing the video sequences with no skips or errors introduced by the testing system, and each system used for testing should be identical. For the test under discussion, the systems have the following specifications:

Monitor: 19" LCD with response time of 12 ms and a native resolution of 1280x1024 non-interlaced.

PC: 3.2GHz dual core processor, 2GB Dual Channel SDRAM, 256 MB PCI Express x16 graphics card, 2x80 GB RAID 0 hard drives at 10,000rpm.

Mplayer[5,6] was chosen as the video player, and ITS developed internal software in Matlab™[7], called NTIA Subjective Tester, to control the test flow. After each 8-second clip is played, the dialog box shown in Figure 3 is displayed for subject scoring (each dialog box displays the number of the current clip).

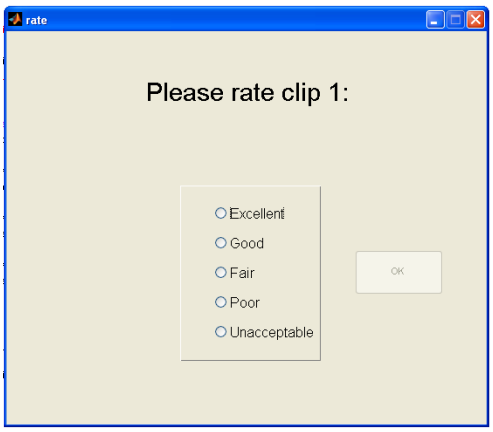

Figure 3. Subjective test scoring dialog box

Once a radio button is selected, the OK button becomes active. After the OK button is selected, the next video clip is played.

ITS is currently executing this subjective test. The results will be reported to VQEG, JRG-MMQA, ITU-T Study Groups 9 and 12, and ITU-R WP6Q.

5. Summary

With the advent of digital transmission of multimedia content, especially over wireless links, methods need to be developed to monitor and measure the quality of the received content. While models and standards exist for objectively measuring the quality of audio and video independently, there is a current need for analogous tools for multimedia – the combination of audio and video. Measurement methodologies need to be adapted from the current standards, and subjective tests need to be conducted in order to develop and train quality assessment models to predict the perceived quality of multimedia signals. Research is currently being conducted to increase the knowledge and standard practice for assessing multimedia quality. These efforts are expected to result in new standards in the ITU to assist industry and users in the measurement of the quality of multimedia applications.

6. References

[1] ANSI T1.801.03-1996, 2003, "American National Standard for Telecommunications - Digital Transport of One-Way Video Signals - Parameters for Objective Performance Assessment."

[2] ANSI T1.518-1998, 2003, "Objective Measurement of Telephone Band Speech Quality Using Measuring Normalizing Blocks (MNBs)"

[5] http://www.mplayerhq.hu/

[6] Certain commercial equipment and materials are identified in this report to specify adequately the technical aspects of the reported results. In no case does such identification imply recommendations or endorsement by the National Telecommunications and Information Administration, nor does it imply that the material or equipment identified is the best available for this purpose.

[7] http://www.mathworks.com/

[3] ITU-R Recommendation BS.1387-1 (2001), "Method for objective measurements of perceived audio quality," Recommendations of the ITU, Radiocommunication Standardization Sector.

[4] ITU-R Recommendation BT.500-11 (2002), "Methodology for the subjective assessment of the quality of television pictures," Recommendations of the ITU, Radiocommunication Standardization Sector.

[5] ITU-R Recommendation BT.1683 (2004), "Objective perceptual video quality measurement techniques for standard definition digital broadcast television in the presence of a full reference," Recommendations of the ITU, Radiocommunication Standardization Sector.

[6] ITU-T COM 12-20 (December 1993), *Experimental combined audio/video subjective test method*, Bellcore.

[7] ITU-T COM 12-37 (September 1994), *Extension of combined audio/video quality model*, Bellcore.

[8] ITU-T COM 12-19 (February 1998), *Relations between audio, video and audiovisual quality*, KPN Research, Netherlands.

[9] ITU-T COM 12-64 (November 1998), *Results of an audiovisual desktop video teleconferencing subjective experiment*, USA.

[10] ITU-T COM 12-61 (November 1998), *Study of the influence of experimental context on the relationship between audio, video and audiovisual subjective quality*, France Telecom/CNET.

[11] ITU-T Recommendation H.261, "Video codec for audiovisual services at p x 64 kbit/sec," Recommendations of the ITU, Telecommunication Standardization Sector.

[12] ITU-T Recommendation J.143 (2000) "User requirements for objective perceptual video quality measurements in digital cable television," Recommendations of the ITU, Telecommunication Standardization Sector.

[13] ITU-T Recommendation J.144 (2004), "Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference," Recommendations of the ITU, Telecommunication Standardization Sector.

[14] ITU-T Recommendation J.148 (2003), "Requirements for an objective perceptual multimedia quality model," Recommendations of the ITU, Telecommunication Standardization Sector.

[15] ITU-T Recommendation P.862 (2001), "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," Recommendations of the ITU, Telecommunication Standardization Sector.

[16] ITU-T Recommendation P.910, "Subjective video quality assessment methods for multimedia applications," Recommendations of the ITU, Telecommunication Standardization Sector.

[17] ITU-T Recommendation P.911, "Subjective audiovisual quality assessment methods for multimedia applications," Recommendations of the ITU, Telecommunication Standardization Sector.

[18] SMPTE 170M, "SMPTE Standard for Television – Composite Analog Video Signal – NTSC for Studio Applications," Society of Motion Picture and Television Engineers, 595 West Hartsdale Avenue, White Plains, NY 10607.

[19] VQEG Multimedia Testplan, http://www.its.bldrdoc.gov/vqeg/projects/multimedia/index.php

[20] A. A. Webster, ed. "Objective perceptual assessment of video quality: full reference television", http://www.itu.int/ITU-T/studygroups/com09/docs/tutorial_opavc.pdf

[21] A. A. Webster, C. T. Jones, M. H. Pinson, S. D. Voran, and S. Wolf, "An objective video quality assessment system based on human perception," *Human Vision, Visual Processing, and Digital Display IV, Proceedings of the SPIE*, Vol. 1913, Feb. 1993, pp. 15-26.

[22] S. Wolf and M. H. Pinson, "Video Quality Measurement Techniques," NTIA Report 02-392, Jun. 2002.

A Network and Data Link Layer Infrastructure Design to Improve QoS in Voice and Video Traffic

J. A. Pérez, V. H. Zárate	J. Jenecek
{ jesus.arturo.perez, vzarate }@itesm.mx	janecek@cs.felk.cvut.cz
Department of Electronics	Department of Computer Science and Engineering
ITESM - Campus Cuernavaca.	Czech Technical University in Prague
Temixco, Morelos, 62589 México	Prague, Czech Republic

Abstract

Currently, there are a lot of e-learning and collaborative platforms to support distance and collaborative learning, however, all of them were designed just like an application without considering the network infrastructure below. Under these circumstances when the platform is installed and runs in a campus, sometimes it has very poor performance. This paper presents a network and data link layer infrastructure design that classifies and prioritizes the voice and video traffic in order to improve the performance and QoS of the collaborative systems applications. This infrastructure has been designed taking in consideration a typical network of a university campus, so that in this way it can be implemented in any campus. After making the design we have made some tests in a laboratory network demonstrating that our design improves 70-130% the performance of these real time collaborative systems which transmit voice and video.

1. Introduction

There are some applications that support collaborative work; however, this does not imply that they are neither effective nor functional. The mayor issues in those applications are focused in the synchronous collaboration because of the problem of managing the information that flows across the network and the mechanism to ensure the quality of the service.

Applications like the "Elluminate Live Academic Edition" provide some tools to work in real-time with other people across the internet; it provides mechanisms to create virtual conferences based in audio and text with pretty good audio quality and also good application performance, even with a low bandwidth connection. It is important to say that this application is not academic and it requires the purchase of a license to allow people to use it. Also we should note, the technology used to ensure the quality of the communication is private [1].

Other applications, like "Synergeia" [2], "Synergo" [3] and "Blackboard" [4], which are designed to operate under the common structure of the internet (client-server), do not provide efficient tools to communicate with other people across the internet in a synchronous manner. The problem in this case is the lack of control and management in the underlying protocols to achieve the demanded Quality of Service (QoS). These issues force this kind of applications to provide only tools that do not exhaust the bandwidth of the communication channel like text based conferences (chat) and shared blackboards or notepads, tools that do not demand a constant flow of information, just chunks of data (messages).

Anyway, even though if the Elluminate software found a way to guarantee some quality in the voice transfer, it does not ensure an efficient management for all types of data. In fact, none of the applications mentioned previously provide tools to interact in a video conference, or to request video on demand. This is because the video information is more sensible to degradation during the transference and consequently, it is more difficult to control the QoS in this kind of data.

All the analyzed collaborative systems work properly when they are exchanging data through a chat or when using off-line communication like e-mail. However, in those which support real time voice and video transmission, the performance and success of the application is conditioned to the performance of the network below. We have discovered that in many cases modern networks are fast enough to support these applications, but the lack of a proper configuration in routers and switches make the applications suffer from performance issues.

Because synchronous communication is the most difficult to implement [5], in order to guarantee the QoS in data transmission independently of the data type we need to provide a designed framework to manage the data flow in the network and data link layer in the OSI Model.

2. Video and Voice requirements

The audio/video information within a videoconference is segmented into chunks by the application, encoded and compressed, put into a series of data packets and sent over the network to the remote end at basically constant intervals. The data packets may arrive at their destination at slightly varying times, and possibly out of order. In order to keep the "real time" impression of an interactive videoconference, the packets must arrive on time and in time to be re-ordered for delivery through the videoconferencing terminal.

Before proceeding, it is important to involve the five fundamental network problems for videoconferencing and for the transmission of voice over IP (VoIP) [6].

1. **Bandwidth** is the fundamental requirement. There must be enough room in a network path for all of the packets to get through unimpeded. This bandwidth need is symmetric: each end requires to transmit and to receive at the same link speed.
2. **Packet loss** is the amount of packets which fail to arrive correctly to their destination. This is due to insufficient bandwidth, transmission errors or high latency. The packet loss percentage must always be below 1% for voice and 2% for video in order to guarantee an understandable communication.
3. **Latency** is the time delay between an event occurring on one site and the remote end seeing it. Latency is introduced both by the encoding/decoding process, and hence depends on the equipment used, and also by the time it takes packets to traverse the network. A disruption in the image can cause a bad playing in the destination, but a disruption in the voice is more important since it makes the transmission not understandable, so it is considered that the biggest latency allowed in the voice transmission to keep a good quality should always be below the 150 ms value.
4. **Jitter** is the average time variation among each received packet. Jitter should always be below 50ms. Multimedia traffic transmission with a great jitter, low bandwidth and great loss could end up in an unacceptable communication [7].
5. **Policies** are introduced by devices like firewalls and network address translation (NAT) controllers that are generally used to hide or protect network elements from the wider Internet.

In this work we are considering that the network of the Autonomous System (AS) where the designed infrastructure is going to be implemented has enough bandwidth for voice and video traffic. We will focus on creating a configuration to minimize the packet loss, latency and jitter for videoconference traffic.

3. General model for traffic prioritization

In order to maintain the high standards that modern applications and collaborative systems require, traffic should always follow a prioritization scheme in order to guarantee specific bandwidth requirements from real time communications, such as voice and video. This scheme can be represented in the form of a general model which applies to all applications which require special conditions (such as maximum delay) to be met. This general model is represented in Figure 1.

After receiving the IP traffic, the first step would be to mark the incoming frames/packets according to our needs. This marking should be done in the way of traffic *classes*. A different class should be specified for every kind of traffic which should be treated in

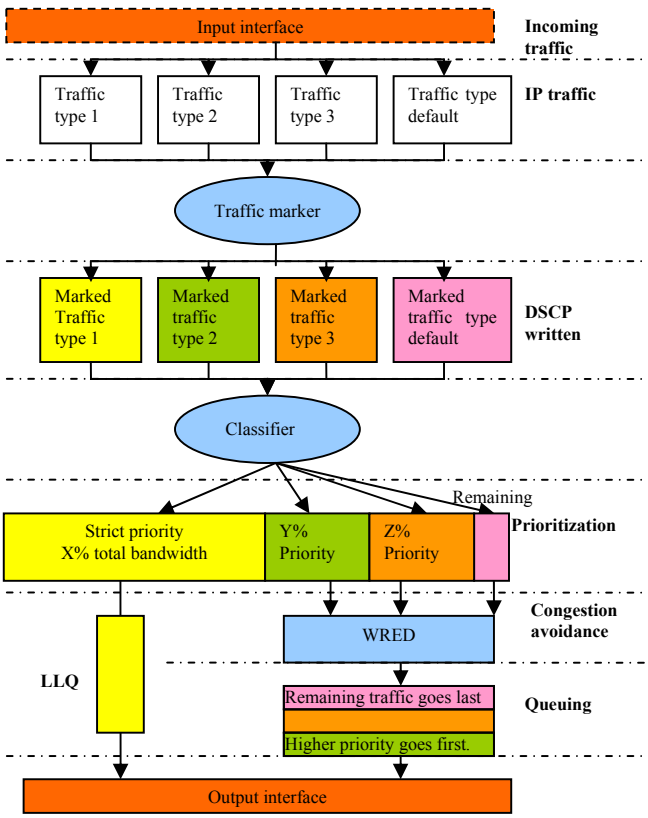

Fig 1. General model for prioritizing network traffic

different way. A common practice is to classify the voice and video traffic in its own class, away from any other type that might cause delays in the processing of the realtime data.

After traffic has been marked, it is ready to be classified according to our own requirements. Since voice and video are the most delay sensitive type of data, they should receive a special treatment in order to

avoid delay at all costs. Any other type would be considered as delay tolerant and so, it would be subject of further processing in order to provide the bandwidth only to those applications that do really need it.

The most important traffic class should receive a *strict priority* using Low Latency Queuing (LLQ). LLQ provides traffic the ability to skip directly to the output interface without having to deal with any congestion avoidance technique, reducing its time to go out from the router. By also specifying a reasonable amount of the total bandwidth, we will be guaranteeing that this type of traffic always has the resources it needs to function properly.

As it is shown in the general model (see figure 1), the rest of the traffic classes must go through WRED congestion avoidance mechanisms and queuing. This process would divide the remaining bandwidth according to the specific policies configured for each data class.

Once all conditions are met and all the policies are applied, the now marked and prioritized traffic is sent through the router's outgoing interface to its destination.

Now that we have shown the general model, we will describe in detail what we propose to improve the performance in each layer of our model.

3.1 Improving data link layer

A shared LAN (using hubs) divides the bandwidth between all the available users, so on average we get much less of the nominal bandwidth, plus increasing the risk of packet loss and jitter due to collisions, while a switched network allows full duplex transmission by using microsegmentation. This totally avoids collisions and provides the highest possible available bandwidth for each of the devices connected to the switch. As additional enhancements, switches open up the possibility of using VLANs for increasing security and reducing broadcast domains, while also allowing the use of trunk interfaces for extending the availability of ports for other medical devices.

In order to improve the performance if the switched network, at layer two, we need to configure in the switches their switch mode operation and traffic prioritization with 802.1p.

3.1.1 Switch mode Operation. How a frame is switched from the source port to its destination is a trade off between latency and reliability. A switch can start to transfer the frame as soon as the destination MAC address is received. This switching method is called cut-through and results in the lowest latency through the switch. However, no error checking is available, but considering the type of application, it is more important to transfer frames faster than to lose some frames. So the switch network infrastructure must support cut-through mode instead of store and forward (or fragment-free modes). The switch cut-through command must be entered for each of the switch's port where cut-through mode should be used.

3.1.2 Traffic Prioritization with 802.1p. When we are using a switched network inside a university campus, sometimes most of the traffic never cross the routers interfaces, if VLANs are used around the campus and the traffic is sent among users of the same VLAN the traffic will never cross the router interfaces, so, there is not a way to prioritize the traffic with layer 3 priorities, for this reason we need to add to our designed infrastructure layer 2 priorities.

The IEEE 802.1p is an extension of the IEEE 802.1Q (VLANs tagging) standard. The 802.1Q standard specifies a tag that appends to an Ethernet MAC frame. The VLAN tag has two parts: The VLAN ID (12-bit) and Prioritization (3-bit), as it is shown in the figure 1. The prioritization field was not defined and used in the 802.1Q VLAN standard. The 802.1P defines this prioritization field.

Using frame tagging as the standard trunking mechanism, as opposed to frame filtering, provides a more scalable solution to VLAN deployment. VLAN frame tagging is an approach that has been specifically developed for switched communications and gives the possibility of using the prioritization field.

The 802.1P standard also offers provisions to filter multicast traffic to ensure it does not proliferate over layer 2-switched networks. The 802.1p header includes a three-bit field for prioritization, which allows packets to be grouped into various traffic classes. It can also be defined as best-effort QoS (Quality of Service) or CoS (Class of Service) at Layer 2 and can be implemented in network adapters and switches without involving any reservation setup. 802.1p traffic is simply classified and sent to the destination; no bandwidth reservations are established.

IEEE 802.1p establishes eight levels of priority. Although network managers must determine actual mappings, IEEE has made broad recommendations. The highest priority is seven, which might go to network-critical traffic such as Routing Information Protocol (RIP) and Open Shortest Path First (OSPF) table updates. Values five and six might be for delay-sensitive applications such as interactive video and voice. Data classes four through one range from controlled-load applications such as streaming multimedia and business-critical traffic - carrying SAP data, for instance - down to "loss eligible" traffic. The zero value is used as a best-effort default, invoked automatically when no other value has been set.

In order to apply the prioritization to be applied throughout the network, switches and NICs that support 802.1p are needed. Considering these priorities we should establish in our switches the five and six value for our traffic in this way, we will enable to prioritize network traffic inside our network campus even though this traffic do not cross router interfaces.

Incoming frames can be examined for a pre-existing priority value, which is then mapped to the 802.1p-specific priority value (according to a matrix provided in the 802.1p specification). The 802.1p priority value can then be assigned to an outbound frame on another medium using this same matrix, providing a standard and topology-independent priority-mapping service. The matrix is used only when a frame is transmitted from one medium or technology to a different one since the priorities are not the same in each medium. In this way also 802.1p allows to provide priority to technologies that for nature did not supported, like Ethernet.

Depending on the switch model, it may be necessary to first activate QoS with the mls qos command. This command is required on both the Catalyst 3550 and the Catalyst 6500. The Catalyst 2950 has QoS enabled by default.

```
switch(config)#mls qos
switch(config-if)#mls qos trust cos
switch(config-if)#mls qos cos 0
```

When configuring the switch you can decide to trust or not in the CoS value of the end device with the configuration above any 802.1Q/p frames that enter the switch port will now have its CoS passed, untouched, through the switch. If an untagged frame arrives at the switch port, the switch will assign a default CoS to the frame before forwarding it. By default untagged frames are assigned a CoS of zero.

In the most of cases it is better not to trust in the CoS value of the end device and force to the switch to assign an specific priority for a specific switch port. Specially it is useful when we know that we will connect a high priority devices to these ports, the better configuration for these ports considering they will received voice traffic is:

```
switch(config)#mls qos
switch(config-if)#mls qos cos 5 override
```

For the rest of the ports which are not being prioritized, the same configuration can be applied, except the CoS value will be reset to the lowest priority, the zero value.

Following these brief rules will create a faster infrastructure in the data link layer to allow faster voice and video transmission,— this is sometimes described as "layer 2 quality of service".

3.2 Improving network layer

When we are willing to provide QoS for traffic that will flow outside of our own LAN, there is the need to specify priorities at layer 3 in order to obtain the desired latency and bandwidth for specific delay sensitive data.

QoS refers to both class of service (CoS) and type of service (ToS). The basic goal of these is to guarantee specific bandwidth and latency for a particular application [8]. To achieve this, we use the Differentiated Services Codepoint (DSCP) field in the packet header to indicate the desired service. This value provides the necessary marking as suggested by the first step of our general model (Figure 1) when dealing with layer 3 traffic.

DSCP redefines the older IPv4 ToS octet and IPv6 traffic class octet. It is composed by the first six bits in the ToS byte, while the IP Precedence value is created with the first three bits in the ToS value. The IP Precedence value is actually part of the IP DSCP value, so both values can not be set simultaneously. If both values are set simultaneously, the DSCP value overwrites the IP precedence one.

The marking of traffic at layers 2 or 3 is crucial to providing QoS within a network, and the decision of whether to mark traffic at any or both of these layers is not trivial. We suggest deciding after the following considerations are made:

- Layer 2 marking can be performed for non IP traffic. This is the only option available for non “IP aware” switches.
- Layer 3 marking will carry the QoS information end-to-end.

We propose to use both DSCP to mark packets through the routed links of the network and also mark the frames using CoS to allow layer 2 devices to provide the QoS requirements of packet at the data link layer.

It is important to mention that a mapping between layer two QoS (CoS) and layer three QoS (DSCP) is possible, as it is presented by Ubik [9] However, since in this paper we are just trying to improve the QoS inside our Autonomous System, we will only propose tools associated with the network edge.

After marking the packages classification will be needed in order to create different classes of traffic with different priority.

3.2.1 Low bandwidth WAN circuits. If any low speed connections exist in the network, and a high portion of the traffic is from the RTP kind, the most proper protocol to use is the Compressed Real Time Transport Protocol cRTP which reduce the consumed bandwidth

by a g.729 voice call since it enables to compress the 40 byte IP/RTP/UDP header to 2 or 4 bytes.

With cRTP the amount of traffic per VoIP call is reduced from 24Kbps to 11.2Kbps, in this way it is possible to double the amount of the calls in one link. In our experiments we will use Cisco equipment, so a basic configuration for CISCO IOS could be:

```
Interface serial 0
   Ip address 192.168.10.1 255.255.255.0
   Ip rtp header-compression
   Encapsulation ppp
```

cRTP is not required to ensure good voice quality. It is a feature that reduces bandwidth consumption. cRTP must be configured on both ends of the link. After Configuring cRTP in the WAN link, if all other conditions are met then the voice quality will be good.

Another important consideration is fragmentation, large packets takes a long time to move across low-bandwidth links (In cases with a WAN link of more than 768 Kbps, the fragmentation feature is not needed), and they can consume the entire VoIP budget. Fragmentation allows to reduce the latency of packets, but the router must also be able to queue based upon fragments or smaller packets instead of the original (prefragmented) packet.

By default with G.729, two 10-ms speech samples are put into one frame. This gives you a packet every 20 ms. This means you need to transmit a VoIP packet out of the router every 20 ms. This means you need to be able to transmit a VoIP packet out of the router every 20 ms.

Blocking directly affects the delay budget, so it is always desirable to keep the blocking delay at 80 percent of your total voice packet size. So in our case we have a 20 ms seconds packet so the maximum blocking delay must be 16 ms. Now, we need to determinate the exact packet fragmentation size for the links we could have in our collaborative environment with the following algorithm:

WAN bandwidth X blocking delay = fragment size in bits

The low bandwidth circuits that we could support are a dial up 56Kbps link or ADSL 256Kbps link, so applying the last algorithm we have:

Fragment size Dial-up link = 56Kbps x 16 ms = 896 bits per second = 112 bytes per second.
Fragment size ADSL link = 256Kbps x 16 ms = 4096 bits per second = 512 bytes per second

As we can see, in the low bandwidth WAN link is it necessary to fragment the packets to 128 or 64 bytes for the dial up connection and to 512 bytes to the ADSL connection.

In order to fragment the packets we can use FRF.12 if we have a frame-relay interface, if we have interfaces that can run PPP, MCML is recommended otherwise we should use IP MTU, even though this last tool can cause many problems since the receiving station's overall performance is affected.

MCML PPP still requires fragments to be classified by IP Precedence, and to be queued by WFQ. A possible configuration could be like this:

```
Router(config)#interface serial 0/0
Router(config-if)#no ip address
Router(config-if)#encapsulation ppp
Router(config-if)#ppp multilink
Router(config-if)#ppp multilink group 1
Router(config-if)#shutdown

Router(config)#interface multilink 1
Router(config-if)#ip address 10.0.100.1
   255.255.255.0
Router(config-if)#fair-queue
Router(config-if)#Ip rtp priority 16384
   16484 50
Router(config-if)#bandwidth 128
Router(config-if)#ppp multilink fragment-
   delay 16
Router(config-if)#ppp multilink
   interleave
```

In this configuration, PPP multilink is configure on the serial 0/0 interface. A new multilink-group 1 interface is then created with IP RTP Priority configured along with MCML PPP and WFQ. Under the serial 0/0 interface, multilink-group 1 maps to interface multilink 1. This enables all the interface multilink 1 attributes to apply to the serial 0/0 interface.

In the Interface Multilink 1 configuration, the fragment-delay of 16 instructs the router to break up any packet into fragments that will not take longer than 16 ms to cross the WAN link. The interleave command will allow PPP to interleave new packets subject to whatever queuing strategy is in place.

For RTP traffic prioritizing at layer 3 over normal bandwidth WAN circuits, our general model proposes the use of Low Latency Queuing (LLQ) to give absolute priority to voice and video traffic over any other traffic over an interface.

3.2.2 Low latency queuing and congestion avoidance techniques. Low latency queuing (LLQ) was designed for being used in realtime applications, such as a videoconference. It brings strict Priority Queuing (PQ) to CBWFQ. Strict PQ allows delay-sensitive data such as voice to be sent directly through the outgoing interface before packets in other queues are sent (as shown in Figure 1). Without LLQ, CBWFQ provides

WFQ based on defined classes with no strict priority queue available for real-time traffic. For CBWFQ, all packets are serviced fairly based on weight and no class of packets may be granted strict priority. This scheme poses problems for voice traffic that is largely delay intolerant, especially delay variation. For voice traffic, variations in delay introduce irregularities of transmission manifesting as jitter in the heard conversation. LLQ provides strict priority queuing for CBWFQ, reducing jitter in voice conversations.

To configure LLQ priority to a class within a policy map, we need to define a class to match the desired traffic. After this, a policy must be created to specify the allowed bandwidth for each class. As an example:

```
R1(config)# class-map match-any
   VoiceTraffic
R1(config-cmap)# match protocol rtp audio
R1(config-cmap)# match protocol rtp video

R1(config)#policy-map StrictPriority
R1(config-pmap)#class VoiceTraffic
R1(config-pmap-c)#strict priority 128
R1(config-pmap-c)#class class-default
R1(config-pmap-c)#fair-queue
```

This would create a class named VoiceTraffic which matches any rtp audio or video packet. Then, the policy would give it a 128kbits bandwidth while also setting the fair-queue as queuing scheme for any other kind of traffic.

When LLQ is not possible to configure, CBWFQ is the best solution, since we can create a specific class and then assign a specific bandwidth that will be enough to guarantee the QoS of the voice traffic. In the following configuration we assign 50% of the bandwidth for the traffic matched by the VoiceTraffic class.

```
Router(config)#policy-map priorityCBWFQ
Router(config-pmap)#class class-default
Router(config-pmap-c)#random-detect
Router(config-pmap-c)#class VoiceTraffic
Router(config-pmap-c)#priority percent 50
```

It is important to show that we also propose (see figure 1) to include a congestion avoidance technique for the rest of the traffic, Weighted Random Early Detection (WRED) with CBWFQ. With the random-detect command we activate WRED, the net result being that the highest priority and lowest bandwidth traffic is preserved, since it starts to drop less important packets once that the net starts to be congested. WRED allows the link to be used more efficiently by selectively dropping packets according to its importance (more packets of lower priority are dropped more than the ones from high priority).

Following the designed rules that we have previously explained in our WAN Links, creating and prioritizing specific classes will have an important performance improvement.

4. Experiments and Results

The aim of this section is to show the performance improvement that a real AS LAN will have after these procedures are followed.

First, we will deploy a network infrastructure using a default configuration (without any kind of priority neither for voice nor video traffic). After that, we will configure the routers and switches with the model that we proposed in the previous section. We will compare results to determine the level of performance improvement obtained with the proposed network design.

The proposed network topology that represent an AS consists on 3 Catalyst 2600 series routers connected through their serial interfaces configured at a 2 Mb/s link speed (simulating an E1 connection). Each of the edge routers will be connected through their fast

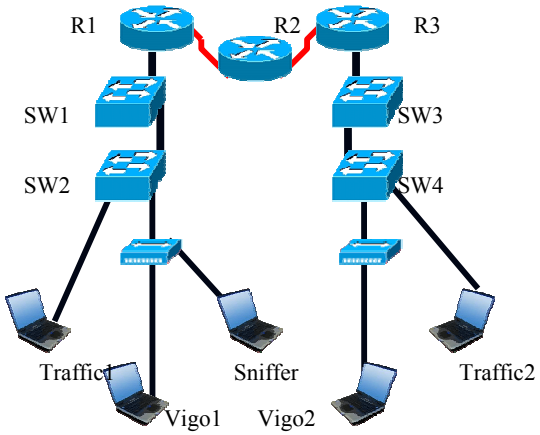

Figure 2. Scenario network topology

ethernet interface with a Catalyst 2900 series switch. Each of these switches connects to one more switch by using its gigabit ethernet trunk interfaces. Finally, the hubs and hosts are connected into these, just as it is pointed out in Figure 2. For each tested scenario we will measure the packet loss, delay and jitter, while testing the data link and network layer.

The routers were configured following a single area OSPF scheme. 802.1q was used on the fast Ethernet interfaces to support the VLAN tagging of the switches

The used IP addresses were as follows:

Used IP Addresses:

R1 Fa0.0 : 192.168.3.1 Trafico1 : 192.168.3.20
R1 S0 : 192.168.1.1 Trafico2 : 192.168.4.20

R2 S1 : 192.168.1.2	Sniffer NIC1 : 192.168.3.10
R2 S0 : 192.168.2.1	Sniffer NIC2 : 192.168.4.10.
R3 Fa0.0 : 192.168.4.1	Vigo1 : 192.168.3.15
R3 S1 : 192.168.2.2	Vigo2 : 192.168.4.15

We will use Vigo videoconference equipment in each end point of the network, while having some other clients generating traffic from protocols like ftp and http. Some other computers will use special software to flood the network with random packets in order to simulate a real scenario. The test for each scenario consists on keeping a videoconference open between two end points of the network while traffic is also being transmitted. We will perform the test of each scenario with and without voice and video priority configurations so that we can measure the improvement percentage.

By recreating a videoconference enabled scenario while also simulating normal network traffic, our testing environment comes very close in terms of reality and thus gives us a much better perception of what would the QoS performance benefit be when applied into a real world case, such as an university campus.

4.1 Endpoints inside the same network – layer 2 priority

The first and simplest scenario describes the typical switched LAN created only by switches. In our simulation, 2 Cisco Catalyst 2950 switches were connected through their Gigabit Ethernet trunk interfaces. For testing, one 3Com 10/100 hub was connected at the Fa0/1 of each of the switches, while also using a traffic generator laptop plugged into the Fa0/2 port of each switch. Both a Vigo videoconference laptop and the sniffer laptop were connected to each of the hubs. The two sniffer cards were inside the same laptop, and each card was connected to a different hub.

A videoconference was established between the 2 Vigo enabled laptops while also injecting traffic from the laptops connected through the Fa0/2 port. All traffic between switches was exchanged through the Gigabit Ethernet trunk interfaces.

The tests ran in our simulated network showed up some slight improvements after applying QoS settings at layer 2. The results were small due to the fact that our layer 2 equipment is able to switch great amounts of data in a very short time, thanks to its fast Ethernet and gigabit Ethernet interfaces. This points out that our attention should be focused into improving the layer 3 prioritization which covers the full AS to where our network is connected.

4.2 Endpoints in different networks – layer 3 priority

In this scenario, the end points are located in different networks, so the traffic will have to go through the router's serial interfaces. In this way we will just evaluate the layer 3 priority. The used network topology can be observed at Figure 2. The sniffer has two cards, each one is connected to a different network.

For this scheme, there are 2 types of router configurations that should be noted. We will refer to them as the *edge routers* and the *middle routers*, being the edge routers the ones that are directly connected to the switches and the middle routers the ones that only use their serial links to communicate the rest of the routers between themselves.

The configuration used for the edge routers were as follows:

```
Router(config)#class-map match-any VOICE-
   VIDEO
Router(config-cmap)# match protocol rtp
   audio
Router(config-cmap)# match protocol rtp
   video
Router(config-cmap)# match protocol rtp
   payload-type "34"
Router(config-cmap)# match ip dscp af41
```

The creation of the VOICE-VIDEO class identifies the RTP traffic commonly used in videoconference. We specify a payload type and a specific dscp value to compare against just to ensure that all our voice/video traffic will be recognized in this class.

```
Router(config-cmap)# class-map match-any
   HTTP-FTP
Router(config-cmap)# match protocol ftp
Router(config-cmap)# match ip dscp af12
Router(config-cmap)# match protocol http
Router(config-cmap)# match ip dscp af11
```

The HTTP-FTP class identifies the traffic we will be using as a 2nd priority. In our tests, our injected traffic is of this kind, so we provide a specific class for it to ensure the router responds as we request.

```
Router(config)# policy-map MARKING
Router(config-pmap)# class VOICE-VIDEO
Router(config-pmap-c)# set dscp af41
```

The MARKING policy is specific of the edge routers. Once in the middle routers, traffic has already been marked so there's no need to do this again.

```
Router(config)# policy-map VOICE-VIDEO
Router(config-pmap)# class VOICE-VIDEO
Router(config-pmap-c)# priority percent
   50
Router(config-pmap-c)# class HTTP-FTP
```

```
Router(config-pmap-c)# bandwidth
   remaining percent 70
Router(config-pmap-c)# class class-
   default
Router(config-pmap-c)# bandwidth
   remaining percent 30
Router(config-pmap-c)# random-detect
```

The VOICE-VIDEO policy gives special treatment to each traffic class specified previously. In the commands entered above, we define a strict 50% traffic priority to all the data matched by our VOICE-VIDEO class. From the remaining bandwidth we chose to give 70% (35% from the total absolute bandwidth) for HTTP and FTP traffic, while giving the rest of the bandwidth to any other kind of traffic not specified in any of our classes.

For each of the edge routers, we applied our policies VOICE-VIDEO and MARKING to the corresponding interfaces. For R1:

```
R1(config)# interface s0/0
R1(config-if)# service-policy output
   VOICE-VIDEO
R1(config-if)# interface fa0/0
R1(config-if)# service-policy input
   MARKING
```

The configurations for the middle router differ from the edge ones, so it doesn't require any marking because R1 and R3 (the edge routers) are doing all the marking themselves. The extra configuration required for the middle router R2 to work was:

```
R2(config)# class-map match-any VOICE-
   VIDEO
R2(config-cmap)# match ip dscp af41
```

The edge routers MARKING policy already set the dscp to af41, so the middle routers can trust this value and only compare the incoming packets against this.

```
R2(config-cmap)# class-map match-any
   HTTP-FTP
R2(config-cmap)# match protocol ftp
R2(config-cmap)# match ip dscp af12
R2(config-cmap)# match protocol http
R2(config-cmap)# match ip dscp af11
```

The same marking was done to the HTTP and FTP traffic, so no need to compare to additional values.

```
R2(config)# policy-map VOICE-VIDEO
R2(config-pmap)# class VOICE-VIDEO
R2(config-pmap-c)# priority percent 50
R2(config-pmap-c)# class HTTP-FTP
R2(config-pmap-c)# bandwidth remaining
   percent 70
R2(config-pmap-c)# class class-default
R2(config-pmap-c)# bandwidth remaining
   percent 30
R2(config-pmap-c)# random-detect
```

This is the same policy specified in the edge routers.

```
R2(config)# interface s0/0
R2(config-if)#service-policy output
   VOICE-VIDEO
R2(config-if)#interface s0/1
R2(config-if)#service-policy output
   VOICE-VIDEO
```

Finally, we apply the policy to our serial interfaces. In contrast to the edge routers, the same service policy needs to be applied to both serial interfaces. Since we don't process nor do any marking from incoming traffic, we do only need to specify the prioritization for the data already marked.

The edge router marks the header and the middle routers are dedicated to give a preferential or deferential treatment to the marked packets with a given DSCP field [10]. By following the previous steps, we will be successfully marking and prioritizing our traffic through all of our routers. It is important to note that the policies must remain equal through all routers to maintain consistency.

After applying this configuration, the sniffer laptop was set to capture and measure the time differences for a 1-way throughput. The following table shows the differences when applying the commands shown above:

Table 1. Experiment results

	Total packets	**Average delay (ms)**	**Jitter (ms)**
Voice (No QoS)	686	27.910	60.870
Voice (QoS)	705	12.036	60.401
Benefit (%)		**131.88**	**.776**
Video (No QoS)	2328	31.209	18.610
Video (QoS)	2399	17.671	17.940
Benefit (%)		**76.61**	**3.60**

During the tests there were no lost packets at all and, as shown, there is a remarkable improvement in both voice and video (131.88% and 76.61% respectively) after applying the QoS settings. However, lets keep in mind that these results were obtained on a simulated network where lots of traffic was being injected into the fast ethernet interfaces to flow through the serial link, thus forcing the router to apply the prioritization. Under higher data load, the benefits margin would have been even bigger.

Conclusions

In this paper we proposed a general guide for enabling QoS inside an autonomous system composed

by several routers and switches in order to provide a more suitable environment for real time traffic used in videoconference. The two created scenarios for simulation of layer 2 and layer 3 infrastructures show up benefits from the implementation of QoS in their policies.

Even though we prioritize the voice and video traffic in our experiments, this model can be applied to any kind of traffic required in collaborative systems.

After running the tests, it's easy to notice the difference between a network with QoS enabled and one without it. The video in both edges appears smoother and the audio is not chopped, no matter what the load in the routers is, as long as the specified priority in the policy maps is enough to handle the video conference demand.

When talking about the urgency to implement QoS at layer 2, we do know that this is not so relevant to keep a good quality conference, since layer 2 only involves devices directly attached into our own switched network, thus providing a connection which depends only on our local hardware, usually fast Ethernet devices. Having a Fast Ethernet switched network provides enough bandwidth for all the devices connected to it, so QoS is not so important as long as the link speed remains constant.

However, when dealing with layer 3, many considerations have to be made since we can not control the traffic coming from other sources. Against this, we must follow the propose model in order to prioritize the outgoing/incoming traffic to be sure that the most important data keeps flowing smoothly without congestions. Inside an AS, this paper provides the required steps to enable QoS in both incoming and outgoing traffic.

The obtained results are a clear sign of the type of improvement which will be obtained in the target AS where these settings are applied (up to 131%). This AS refers to the final network where our collaborative system could be connected, enhancing the quality of their communications while allowing for total control of the traffic flowing through it.

By using a scenario recreating real traffic with the use of http, ftp, pings and udp traffic, our tests come close to reality, showing up that our general model can be successfully applied into a real world scenario while obtaining benefits close to the ones that we got here.

With this general configuration model, we can guarantee an optimal performance inside the AS, translating into a direct benefit to the network where the collaborative systems are set down.

References

[1] Elluminate, I. "Elluminate Live, Academic Edition". Internet Web Site. http://www.elluminate.com. Consultation date: October 2004.

[2] ECOLE. "Synergeia". Internet Web Site. http://www.ecolenet.nl/best/synergeia.htm. Consultation date: October 2004.

[3] Patras, U. "Synergo". Internet Web Site. http://www.ee.upatras.gr/hci/synergo/. Consultation date: October 2004.

[4] Blackboard, I. "Blackboard Portal System". Internet Web Site. http://www.blackboard.com. Consultation date: October 2004.

[5] Avouris, N.; Margaritis, M.; Komis, V. "Real-Time Peer Collaboration In Open And Distance Learning". Proceedings: 6th Hellenic European Conference On Computer Mathematics & Its Applications. Athenas. September 2003.

[6] S. Vegesna, "*IP Quality of Service*", Cisco Press, 2001. ISBN 1-57870-116-3.

[7] Monsour Yishay, Patt-Shamir Boaz, "Jitter control in QoS networks", IEEE/ACM Transactions on Networking (TON), Vol. 9, Issue 4, August 2001.

[8] L. Burgstahler *et al*, "Beyond Technology: The Missing Pieces for QoS Sucess", *Proceedings of the ACM SIGCOMM 2003 Workshops*, Aug 2003, pp: 121-130.

[9] Seven Ubik, Josef Vojtech. *QoS in Layer 2 Networks with Cisco Catalyst 3350*. **CESNET Technical Report 3/2003**.

[10] Fineberg, V., "A practical architecture for implementing end-to-end QoS in an IP network", IEEE Communications Magazine, Vol. 40 , Issue: 1, Jan.2002, pp.122– 130.

A Study on the Acoustic Properties of the Constituent Films in Solidly Mounted Resonators Using Picosecond Ultrasonic Waves

Ta-Ching Li[a,*], Shih-Piao Yu[b] , Le-Ye Tang[c] , Yong-Kang Hsiao[d], Nen-Wen Pu[d], Ben-Je Lwo[e], and Chin-Hsing Kao[b]

[a]School of Defense Science, [b]Department of Weapon System, [c]Department of Electrical Engineering, [d]Department of Applied Physics, and [e]Department of Mechanical Engineering,
Chung-Cheng Institute of Technology, National Defense University, Dahshi, Taoyuan 335, Taiwan.

*Corresponding author E-mail address: g950102@ccit.edu.tw; Tel: +886-3-3900119; Fax: +886-3-3900119

Abstract— Structurally robust, small sized solidly mounted resonators are suitable for mobile and wireless communication system. The aim of this paper is to investigate the acoustic properties of the constituent films in solidly mounted resonators through picosecond ultrasonic waves. These films include molybdenum, tungsten and silicon dioxide; they are commonly employed as constituent layers in Bragg reflectors. In addition, to decrease the parasitic capacitance in resonators, resistivity-adjustable TaN_x films were chosen as a high-acoustic-impedance layer to replace the conventional metallic films, i.e., molybdenum, and their electrical and acoustic properties were studied in detail. The acoustic impedance and velocity of reactively sputtered TaN_x films were measured from the reflectivity response of picosecond ultrasonic waves, and the value of acoustic impedance is higher than molybdenum for films grown at low N_2 concentration.

I. Introduction

As wireless networks grow rapidly in the spectrum from 500 MHz to 10 GHz and the device sizes shrink continuously, the need for integration of the front end of radio frequency becomes more apparent. Great efforts are made to develop miniature filters as spectrum crowding gets worse. One of the most promising filters for the integration is the film bulk acoustic resonators (FBARs) [1,2]. Recently, a new type of solidly mounted resonators (SMRs), which are constructed in a different way from the membrane-type FBARs, have been developed to achieve enhanced robustness. A Bragg mirror made up of a stack of alternate high- and low-acoustic-impedance (Z) quarter-wave thin films replaces the free surface at the bottom of the membrane-type FBARs. The Bragg mirror is solidly mounted on the substrate, and is used for acoustical isolation from the lossy substrate so that a high-Q resonator can be obtained. The required layer thickness and the resulting acoustic reflectivity are strongly dependent on the sound velocity and the acoustic impedance of the film. In this study, we use the high spatial resolution acoustic pulses, i.e., picosecond ultrasonic waves, to measure the velocities of various high- and low-acoustic-impedance films, and to estimate the acoustic impedance of thses films. Furthermore, since Bragg reflector is a natural MOS capacitor with a metal/oxide/substrate structure, lower parasitic capacitance can be achieved by replacing metallic films with high-resistivity films. Thus we also try to investigate the electric resistivity of the films. These results are useful for the researchers and designers of SMRs.

Tantalum nitride (TaN_x) films have received considerable interest in recent years owing to their unique properties including thermal stability [3-5], good optical properties [6], high conductivity [7-10], corrosive resistivity [11,12], and hardness [13,14], making it very attractive for use as structural elements in integrated circuits. Due to their high mass density (therefore high acoustic impedance) and adjustable resistivity (film resistor), TaN_x films are considered a replacer for the present metallic high-acoustic-impedance layers in acoustic wave resonators to overcome the parasitic capacitance problem [15].

Picosecond ultrasonics has proven useful for studying acoustic properties of thin films [16,17]. This technique measures the optical transient reflection change (ΔR) induced by acoustic waves through a pump-probe detection technique. The acoustic waves are excited with a “pump” light pulse generated by an

ultrafast laser, and detected by a delayed "probe" light pulse derived from the same laser. The acoustic velocities are next obtained by measuring the echo times of the elastic strain waves (i.e. acoustic waves) reflected from the interfaces or the reflective substrate and re-enter the surface of the transducer film [18,19].

II. Experiments

All test films were deposited on p-type (100) Si substrates using radio frequency (rf) magnetron sputtering from a 50.8-mm-diameter target (99.95%). In addition, to study the adjustable resistivity in TaN films, reactive sputtering was employed. The substrate, which was neither cooled nor heated externally, was kept 60 mm from the target holder. The background pressure was vacuumed initially to $2 \sim 3 \times 10^{-6}$ torr. The sputtering parameters of deposited films are listed in table I.

The cross-sections of the films were characterized with a scanning electron microscope (SEM, Jeol, model 6340F) with an operating voltage of 5 kV. The film resistivity was calculated from the sheet resistance measured with a four-point probe and the film thickness measured with a surface profilometer (Veeco, model Dektak III). The densities (D) of TaN films were calculated from the formula $D = M/V$, where the mass M was measured by a micro-balance and the volume V was obtained from the fixed area (600 ± 50 mm^2) multiplied by the film thickness.

To optically excite the ultrasonic waves, a self-made mode-locked Ti: Sapphire laser with 50-fs pulsewidth, 800-nm wavelength, and 86-MHz repetition rate was used. The average output power of the laser is about 200 mW, which corresponds to an energy of 2.3 nJ per pulse. The experimental setup with which we made our measurements has been described in detail previously [20].

A very thin Ni layer was deposited on these samples as the light-to-sound transducer. In addition, to ensure the accuracy of acoustic velocity measurement, the area for thickness measurement and the laser probed spot on the sample should be as close as possible. To this end, a mark was first made on the substrate, and we measured both the thickness and the acoustic response in proximity to this mark point.

III. Results and Discussion

In the following plots of reflectivity changes (ΔR) vs. delay time, we use τ to denote the round trip time of acoustic waves, and the velocity of the test film can be calculated from the echo time τ and the film thickness. In addition, the amplitude of ΔR reveals the ratio of high to low acoustic impedances, where acoustic impedance (Z) is defined as the product of density and sound velocity. The results of all test films are described as follows:

Table I. The sputtering parameters for test films.

Film	Pressure (mtorr)	N_2 partial pressure ratio (%)	Thickness (nm)
Mo[a]	7	-	177 ± 5
W[a]	7	-	100 ± 5
SiO_2[a]	7	-	200 ± 5
TaN[b]	3	0	246 ± 5
TaN[b]	3	3.3	200 ± 5
TaN[b]	3	8.5	290 ± 5
TaN[b]	3	25	188 ± 5

a: sputtered with an rf power of 80 W

b: reactively sputtered with an rf power of 200 W

A. The Mo film

Figure 1 shows the results of ΔR vs. delay time for the Mo film. The structure of test sample is depicted in the inset of the figure. A SiO_2 film (lower Z) is used as a reflective layer. Owing to the photoelasticity in Mo, the film itself serves as the light-to-sound transducer. It is obvious that the delay time τ_2 is twice of τ_1, and the amplitude of ΔR at τ_1 is bigger than at τ_2. These are the responses of the successive acoustic echoes bounced back and forth inside the film. According to the thickness of Mo in Table I and the round trip time τ_1, the velocity of the Mo film is about 7300 ± 50 m/s.

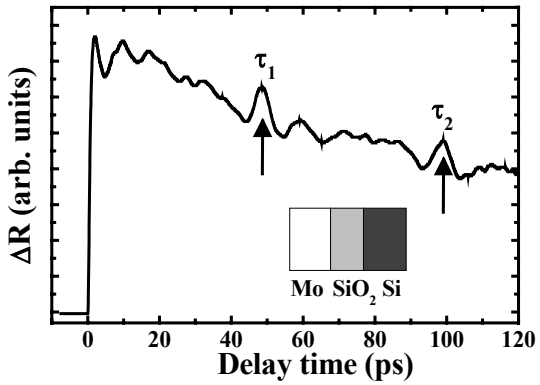

Fig. 1 $\Delta R(t)$ of Mo deposited on SiO_2. The inset is not drawn to scale.

B. The W film

Figure 2 shows the results for Ni/W film. (Ni has higher piezoreflectance [21], so the response ΔR can be easily observed.) The acoustic wave was generated by the Ni transducer. The first response at τ_1 is due to the echo reflected from the Ni/W interface, which shows no phase change because $Z_{Ni} < Z_W$ On the other hand, the second response at τ_2 is caused by the echo reflected from the W/SiO_2 interface, which does show a reflective phase shift because $Z_W > Z_{SiO_2}$. (Therefore, the shape of the response is reversed). According to the thicknesses of Ni (127 nm) and W in Table I, and the round trip times τ_1 and τ_2, the velocities of Ni and W films are about 5450 ± 50 m/s and 6230 ± 50 m/s, respectively.

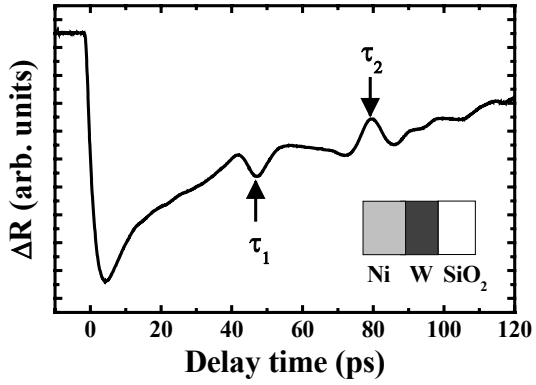

Fig. 2 $\Delta R(t)$ of the Ni/W/SiO_2 film. The thickness of Ni is 127 nm. The inset is not drawn to scale.

C. The SiO_2 film

Figure 3 shows the results of $\Delta R(t)$ for the Mo/SiO_2/W structure. Since SiO_2 has no photoelastic response, a thin Mo layer (about 20 nm) was deposited on SiO_2 film as a transducer. The plot shows two echoes: the one that appears first is smaller and denoted as τ', the other is bigger and denoted as τ_1. This phenomenon is a double-transducer effect [22]. The echoes at τ' and τ_1 are induced by W (underneath SiO_2) and Mo, respectively. This is because considerable amount of the pumping laser light transmits through the thin Mo transducer and the transparent SiO_2, and reaches the W layer. So an extra acoustic wave was generated by W. This is the reason why the delay time τ' is half of τ_1. According to the thickness of SiO_2 and the delay times, the velocity of SiO_2 is about 6450 ± 50 m/s.

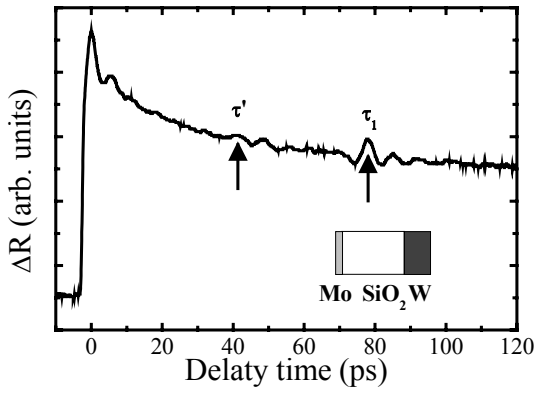

Fig. 3 $\Delta R(t)$ for the Mo/SiO_2/W structure. The inset is not drawn to scale.

D. The TaN_x films

The resistivity of TaN_x films can be changed through reactive sputtering process. Figure 4 shows the average electrical resistivities ($\bar{\rho}$) of the as-deposited TaN_x films as a function of P_{N_2}. The resistivity rises steeply (by 6 orders of magnitude) with P_{N_2}.

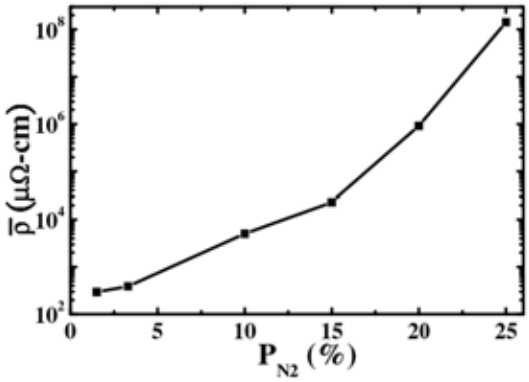

Fig. 4 The average resistivity ($\bar{\rho}$) of the as-deposited TaN_x films as a function of P_{N_2}.

The SEM pictures, which show the various microstructures in TaN_x films except for 0% P_{N_2}, are displayed in Fig. 5. The four nitrogen partial pressure ratios, 0%, 3.3%, 8.5%, and 25%, were chosen because they correspond to four different microstructures, i.e., pure Ta (0%), poly-crystal (3.3%), double-structure (8.5%), and amorphous structure (25%), respectively. In these cross-sectional pictures, an amorphous-structure sublayer sandwiched between the Si substrate and a columnar-structure sublayer can be recognized for 8.5% P_{N_2} (Fig. 5(b)). The thickness of the void-filled amorphous sublayer is an important factor influencing the mechanical and electrical

properties. Our previous study [23] found that it has strong dependence on P_{N_2}: as P_{N_2} increases from 3.3%, the thickness of this amorphous part increases gradually, and eventually develops throughout the film.

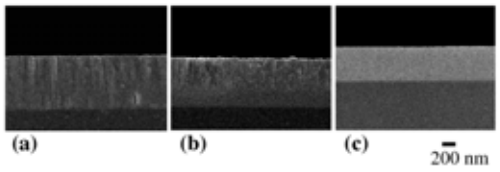

Fig. 5 SEM pictures of TaN_x films deposited at P_{N_2} of (a) 3.3%, (b) 8.5%, and (c) 25%.

The density of TaN_x films was found to decrease with increasing P_{N_2} as shown in Fig. 6. This is because the void-filled amorphous structure is more porous than the poly-crystalline structure. The acoustic impedance of different TaN_x films is also influenced by the various densities.

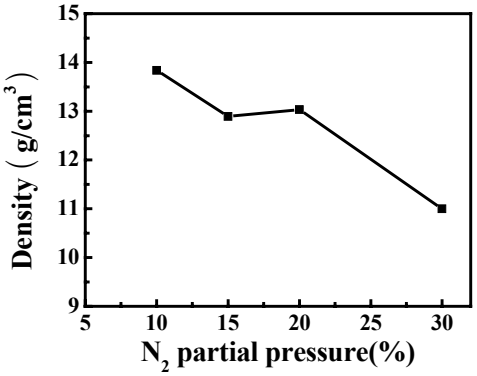

Fig. 6 The density of the as-deposited TaN_x films as a function of P_{N_2}.

To study the effects of microstructure on acoustic impedance (Z) and velocity, the above-mentioned four different structures were investigated with picosecond ultrasonics. Figure 7 shows the results of ΔR vs. delay time for these structures. Among these results, Fig. 7 (a) has the largest magnitude of the echo signal. This is because of the largest mismatch of Z between Ta and Si substrate. Having the same Si substrate, the Z of TaN_x can be estimated from the magnitude and polarity of the acoustic echo. Evidently, the acoustic impedance Z decreases with P_{N_2}.

In Fig. 7 (c), the shape of the echo signal pointed by the dotted arrow is broader than the others; moreover, its echo time (τ') is much shorter than those of the other samples (τ_1). In light of the double structures (for 8.5% P_{N_2}) revealed by SEM pictures, this broader echo is reflected from the polycrystalline/amorphous interface inside the TaN_x film. (The echo reflected from the film/substrate interface is too weak to be detected.) The broader shape was caused by the roughness of this interface. Judging from the polarity of the echo signal in Fig. 7(c), the Z of a polycrystalline structure is larger than that of an amorphous one. Calculated from the measured film thicknesses and the echo times in Fig. 7 (a), (b), and (d), the acoustic velocities for 0%, 3.3%, and 25% P_{N_2} films are $\mathbf{5000 \pm 50}$, $\mathbf{5300 \pm 50}$, and $4600 \pm 50 m/s$, respectively.

From the above data of density and velocity in TaN_x films, we anticipate the acoustic impedance ratio of Mo to TaN_x for 3.3% P_{N_2} to be about 0.98, i.e., the value of Z_{TaN} is larger than Z_{Mo} at low P_{N_2}. Thus we conclude that TaN_x is a promising high acoustic impedance material which can replace the present metallic films.

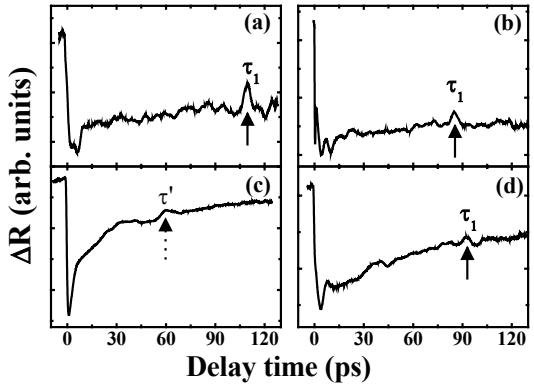

Fig. 7 The reflectance change $\Delta R(t)$ for TaN_x films deposited under (a)0% (b)3.3% (c)8.5% and (d)25% P_{N_2}. τ_1 is the round trip time, and the values are (a)109 ps (b)85 ps (c)unknown and (d) 92 ps, respectively. In (c), τ' is the echo time for the acoustic wave reflected from the polycrystalline/amorphous interface inside the TaN_x film.

IV. Conclusion

In this study, the acoustic properties of the constituent films in solid mounted resonators, which are the most advanced technology for high frequency in the next generation, were investigated using picosecond ultrasonics technique. For these films having different optical properties, different structural designs of test films are required to enhance the signal of reflectivity changes. These results are useful for the designers to improve the performance of solid mounted resonators. In addition, to reduce the parasitic capacity of filters made by SMR, the properties of resistivity-adjustable TaN_x films were studied. We found that the TaN_x films have higher acoustic impedance than Mo at low nitrogen partial pressure ratio. We believe that the simultaneous high electrical resistivity and high acoustic impedance of TaN_x films will be useful for minimizing the parasitic capacitance in acoustic wave devices.

Reference

[1] K. M. Lakin, G. R. Kline, K. T. McCarron, IEEE Trans. Microwave Theory Tech. 41 (12) (1993) 2139.

[2] R. Weigel, D. P. Morgan, J. M. Owens, A. Ballato, K. M. Lakin, K. Hashimoto, C. C. W. Ruppel, IEEE Trans. Microwave Theory Tech. 50 (3) (2002) 738.

[3] J.O. Olowolafe, C.J. Mogab, R.B. Gregory, M. Kottke, J. Appl. Phys. 72 (1992) 4099.

[4] T. Oku, E. Kawakami, M. Uekubo, K. Takahiro, S. Yamaguchi, M. Murakami, Appl. Surf. Sci. 99 (1996) 265.

[5] M.H. Tsai, S.C. Sun, C.E. Sun, C.E. Tsai, S.H. Shuang, H.T. Chiu, J. Appl. Phys. 79 (1996) 6932.

[6] S.M. Aouadi, and M. Debessai, J. Vac. Sci. Technol. A 22 (2004) 1975.

[7] X. Sun, E. Kolawa, J-S. Chen, J. S. Reid, and M-A. Nicolet, Thin Solid Films 236 (1993) 347.

[8] B. Mehrotra, and Jim Stimmel, J. Vac. Sci. Technol. B 5 (1987) 1736.

[9] T. Riekkinen, J. Molarius, T. Laurila, A. Nurmela, I. Suni, and J.K. Kivilahti, Microelectronic Eng. 64 (2002) 289.

[10] Y.M. Lu, R.J. Weng, W.S. Hwang, and Y.S. Yang, Thin Solid Films 398-399 (2001) 356.

[11] M. Nordin, M. Larsson, and S. Hogmark, Surf. Coat. Technol. 120-121 (1999) 528.

[12] S.K. Kim and B.C. Cha, Thin Solid Films 475 (2005) 202.

[13] R. Westergard, M. Bromark, M. Larsson, P. Hedenqvist, and S. Hogmark, Surf. Coat. Technol. 97 (1997) 779.

[14] R. Saha and J.A. Barnard, J. Crystal Growth 174 (1997) 495.

[15] K.M. Lakin, G.R. Kline, and K.T. McCarron, IEEE Trans. Microwave Theory and Technol. 43 (1995) 2933.

[16] H.T. Grahn, H.J. Maris, J. Tauc, IEEE J. Quantum Electron. 25 (1989) 2562.

[17] D.M. Profunser, J. Vollmann, J. Dual, IEEE Ultrason. Symp. (2002) 995.

[18] H.T. Grahn, H.J. Maris, J. Tauc, Appl. Phys. Lett. 53 (23) (1988) 2281.

[19] C. Thomsen, H. T. Grahn, H. J. Maris, J. Tauc, Phys. Rev. B 34 (6) (1986) 4129.

[20] T.-C. Li, N.-W. Pu, B.-J. Lwo, L.-J. Ho, and C.-H. Kao, Thin Solid Films (in press, 2005).

[21] G. L. Eesley, B. M. Clemens, and C. A. Paddock, Appl. Phys. Lett. **50**, 717 (1987).

[22] T.-C. Li, N.-W. Pu, B.-J. Lwo, E-Y Pan and C.-H. Kao, Appl. Phys. Lett. (no published, 2005).

[23] T.-C. Li, B.-J. Lwo, N.-W. Pu, Shih-Piao Yu, and C.-H. Kao, Surf. Coat. Technol. (no published, 2005).

On Controlled Node Mobility in Delay-Tolerant Networks of Unmanned Aerial Vehicles

Daniel Henkel, Timothy X Brown
University of Colorado at Boulder
530 UCB, ECOT 311, Boulder, CO, 80303
Tel.: 303-492-1630, Fax: 303-492-1112
{henk,timxb}@colorado.edu

Abstract

This paper explores the use of controlled node mobility in delay-tolerant networks. Mobile nodes are used to ferry data between otherwise unconnected network nodes. In particular, the problem of efficient route design for aerial data-ferrying nodes is studied. We introduce two new approaches to route design, namely the chain-relay model and the conveyor-belt model. Analytical evaluation of the proposed route designs shows their relative merits depending on node velocity, data rate, and buffer size. Further, we introduce a new problem framework, which departs from the traditional approach of seeing communication as limited to one data rate within a circular region around a node. Our model takes into account multiple data rates varying with the separation distance from a node. Preliminary evaluation suggests performance improvements over existing approaches when combining our route designs with this problem framework.

1 Introduction

Mobile ad-hoc networks (MANETS) are expected to play an increasing role in making ubiquitous computing and communications become a reality. Already mesh networks are increasingly being deployed to give people access to the Internet where no wired infrastructure is readily available. However, with the advent of ever cheaper computing power in ever smaller mobile devices the potential of MANETS goes beyond accessing the Internet. Example applications include search and rescue coordination without fixed infrastructure; scientific observation in high-risk areas; and inter-vehicle communication for collision avoidance.

The data that is used by applications in these networks can be classified as either *real-time* or *delay-tolerant*. Real-time traffic includes voice over IP, video conferencing, or monitoring of critical processes. Delay-tolerant traffic does not carry an urgency and only eventual, reliable reception is important. Examples are email, instant messages, or sensor data of a long-term experiment.

For a wireless ad-hoc network to be able to carry real-time data, there needs to be *contemporaneous, reliable path* from sender to receiver. Such a path is more likely the shorter the sender and receiver separation and the more dense the radio nodes.

Real-time communication is challenging for wireless mobile ad-hoc networks because of inherent variability in wireless connections combined with changing node positions. In the extreme case of sparsely connected, moving nodes, some nodes might not be able to connect with any other node for a long time. In this scenario only delay-tolerant communication is feasible.

The Ad-hoc UAV-Ground Network project (AUGNet) at the University of Colorado lends itself to studying methods to support sparsely-connected, delay-tolerant networks [1, 2]. AUGNet features fast-moving planes which can span large distances in a relatively short time. It is exploring how moving nodes can be used to reduce delay and improve reliability of communication in sparsely connected networks. Controlled plane mobility can increase battery life of ground nodes, enable coordination of large-scale deployments of and among swarming planes, and simplify data collection and distribution among ground nodes.

This paper investigates the role of ferrying in two communication models: the first has a constant data rate up to a maximum distance and zero rate otherwise; the second has a continuously variable rate that decreases with distance and is always non-zero. Several models of ferry-

ing are developed for both a single transfer between two nodes and a transfer between several nodes and a hub. The results are three-fold. First, they reveal which fixed-rate ferrying model is the most suited for a given set of external operating conditions. Second, a delay-optimal scheduling algorithm is derived which assigns ferry time to nodes in a star network. Third, network performance is greatly improved by considering a variable data rate model where the data rate is a continuous function of node separation distance.

The paper proceeds as follows: Section 2 describes our model for delay-tolerant networking and route planning. Section 3 analyzes route design under full knowledge and introduces our new problem framework. Section 4 gives an overview of related work, and Section 5 concludes the paper with an outlook on future work.

2 Modeling Delay-tolerant Networks

2.1 System Components

Task Nodes
Task nodes form the basic sparsely connected ground network. The nodes operate over an area A, called the *survey area*. For our purposes these nodes are assumed to be stationary. Every node has a buffer of size b_{node} bytes and a wireless radio capable of transmitting or receiving at data rate R_{node} in a coverage area forming a concentric sphere around the node with radius r_{node}. For this scenario the data rate is considered constant within the coverage area.

It is assumed that the separation distance l between task nodes is greater than the communication range of the radios, i.e., $l > r_{node}$. Thus there is no connection between nodes and no possibility for routing traffic among them without the help of some other nodes.

A node can send data to a receiver that is at a distance less than r_{node} as long as the receiver is outside of the interference range of any other transmitting node. The interference range is assumed equal to the communication range.

We refer to this model as the fixed radius communication model.

Ferries
Mobile nodes called *transport ferries* are capable of moving around the survey area to any location requested. There are n ferries on the survey area at any given point in time, denoted $F_1, F_2, \ldots, F_n$. They are areal vehicles and the motion can be controlled arbitrarily. The motion is assumed linear with speed v, but ferries can also stop, e.g., for data exchange. Every ferry is equipped with a radio capable of transmitting or receiving at data rate R_{ferry} in a coverage area forming a concentric sphere around the ferry with radius r_{ferry}. The ferry can buffer b_{ferry} bytes of data. For simplicity we are assuming buffer size, communication radius, and data rate are the same for both, task nodes and ferries and so the subscripts will be dropped. Ferries can talk to nodes or other ferries, but not at the same time.

2.2 Performance Metrics

The analytical evaluation of the proposed ferry path models will compare the following performance measures. They are computed for the ideal case of perfect coordination between ferries.

2.2.1 Throughput

The throughput as experienced by the sender or receiver can be calculated as follows.

$$T = \frac{\text{buffers delivered per cycle}}{\text{cycle period}} \tag{1}$$

Herein the cycle period comprises the total time for a ferry to follow a given path and return to the original position and state. It includes ferry travel times plus times to fill and empty buffers. Such a cycle may involve one or more buffers of traffic delivered.

2.2.2 Average Packet Delay

The average delay experienced by one packet can be derived by looking at the best-case and worst-case delays incurred by the system with subsequent averaging[1]. The best-case delay occurs when a packet stored in the source buffer gets loaded onto the ferry last, i.e., just before departure. The worst-case delay occurs when a packet arrives at the source buffer just as the ferry departs, i.e., it gets buffered at the source as the first packet until arrival of the next ferry. Figure 1 illustrates this method.

2.2.3 Buffers

Packet delay includes the ferry travel times, time that a packet spends waiting in ferry buffer while other packets load and unload, and time for buffering at the source. Propagation delay is assumed negligible. Packets are considered small relative to buffer sizes so that buffers

[1]The validity of this approach can be shown for the chain relay and conveyor belt models. For other models it is an approximation.

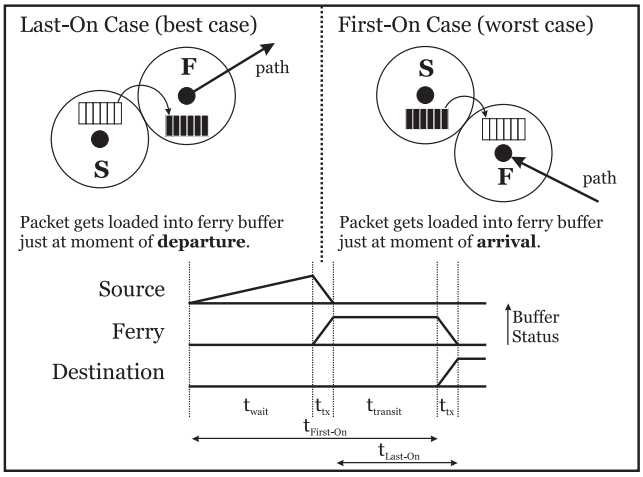

Figure 1: The two boundary cases of packet delay and corresponding timing diagram

smoothly fill and empty. Traffic flows are also smooth and continuous. Thus data is modeled as a fluid for simplicity so that we can focus on the role of mobility.

We assume a FIFO buffer policy. If R is the transmit rate then $\frac{b}{R}$ is the time required to completely fill or empty the buffer.

2.3 Traffic Flow and Ferry Path Models

When looking at traffic flow, there can be three different models: (a) Star model: all traffic flows between one special node (the hub) and the k task nodes, (b) Uni-directional model: traffic can flow between any nodes, but only in one direction, and (c) Bi-directional model: traffic can flow between all nodes in both directions.

We evaluated two different path models for ferry movement, namely the *chain-relay model* and *conveyor-belt model*. The following sections introduce these mobility models in the special case of $k = 1$ nodes communicating with the hub uni-directionally.

2.3.1 Chain-Relay Model

In this model it is assumed that all participating ferries are distributed along a connecting path from source to destination, forming a chain. The source passes available data on to the first ferry, which physically transports it some way along the path, hands it off to the next ferry, and returns to the source. This hand-off procedure is repeated until the last ferry hands the data to the destination. Figure 2 depicts this model.

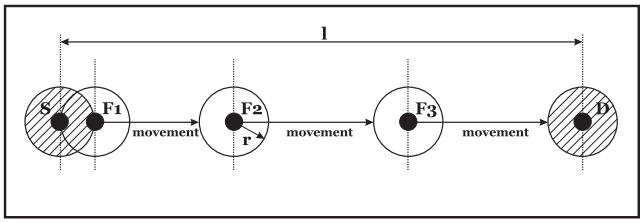

Figure 2: The Chain-Relay Model

Throughput and Ferry Travel Space
The cycle time is the time for F_1 to fill its buffer, travel to within the range of F_2, offload the buffer, and travel back to the source node. The traveling distance is the total separation distance from sender to receiver, l, less $n + 1$ times the communication radius, divided by the number of ferries that operate on the link:

$$d_{neighbor}(n) = \frac{l - (n+1)r}{n}.$$

Therefore

$$t_{cycle} = t_{travel} + t_{tx} = 2\frac{l - (n+1)r}{n\,v} + 2\frac{b}{R}$$

and from (1) the throughput can be stated as

$$T = \frac{b}{2\left(\frac{l-(n+1)r}{n\,v} + \frac{b}{R}\right)}. \tag{2}$$

In this model even if $v \to \infty$ the maximum throughput is $\frac{R}{2}$. Substituting $T = \frac{R}{2}$ into (2) and solving for n yields $n = \frac{l}{R} - 1$. In this case, $d_{neighbor} = 0$, and we reach the "saturation" case of a connected chain of relays[2].

Delay Calculation
We compute the packet delay by averaging over worst possible delay and best possible delay. (a)The First-on

Case (worst-case delay)
The packets in the chain model get handed off between neighbor-ferries and one cycle happens between two neighbors. This cycle time needs to be accounted for in the wait time for the packets. The total delay seen by a single packet from source to destination consists of the following times.

[2]Note that when $d_{neighbor} = 0$ we rely on the interference distance to be equal to the communication distance. In fact, the nodes move a small distance away from the sending node and out of its interference range before communicating with the next node in the relay. A longer interference range would require a longer minimum separation between nodes.

i) *packet wait delay* until neighbor ferry shows up again and stops at distance r from center of source; this includes actual travel time of ferry as well as off-loading and loading at remote neighbor, plus off-loading time at the source. This is not the average wait time since the ferry had just departed and a full cycle needs to be traveled.

$$t_{wait} = 2\frac{l-(n+1)r}{nv} + \frac{b}{R}$$

ii) *ferry loading time.* Even though the first packet is loaded immediately into the ferry buffer, the packet still has to wait until all packets are loaded and the ferry is ready to depart.

$$t_{loading} = \frac{b}{R}$$

iii) *ferry travel delays* to n remote neighbors, including the destination.

$$t_{transit} = n\frac{l-(n+1)r}{nv}$$

iv) *buffer exchange time* with $n-1$ neighbors. We assume an empty neighbor buffer.

$$t_{exchange} = (n-1)\frac{b}{R}$$

(b)The Last-on Case (best-case delay) A packet experiences the following delays from source to destination.

i) n transit delays between neighbors.

$$t_{transit} = n\frac{l-(n+1)r}{nv}$$

ii) n buffer exchanges between neighbors, including the destination.

$$t_{exchange} = n\frac{b}{R}$$

The average delay of packets from source to destination is the average of best and worst case delays:

$$delay^{chain} = \left(1+\frac{1}{n}\right)\frac{l-(n+1)r}{v} + \left(n+\frac{1}{2}\right)\frac{b}{R} \quad (3)$$

2.3.2 Conveyor-Belt Model

In this model it is assumed that every ferry travels the complete distance from source to destination, then returns to the source. Multiple ferries are sharing the same path, which increases throughput and robustness of the system. Figure 3 depicts the model.

Throughput and Buffer Status

One cycle in the conveyor belt model is given by a complete round-trip from source to destination including buffer transmission times. The corresponding cycle time is thus given by

$$t_{cycle} = t_{travel} + t_{tx} = 2\frac{l-2r}{v} + 2\frac{b}{R}$$

Since there are n ferries hauling data between two nodes, the data delivered in one cycle time is $n \cdot b$. According to (1), the total throughput is

$$T = \frac{n \cdot b}{2\frac{l-2r}{v} + 2\frac{b}{R}} = \frac{n}{\frac{2(l-2r)}{v \cdot b} + \frac{2}{R}}. \quad (4)$$

R is the maximum rate at which a node can transmit. $T = R$ implies that for the node to be transmitting at its maximum data, at least n_{opt} ferries are necessary:

$$n_{opt} = \left\lceil 2\left(1+\frac{l-2r}{v}\frac{R}{b}\right)\right\rceil.$$

When $n > n_{opt}$, ferries arrive faster than buffers can fill. In this case, either (a) ferries get queued at the source or (b) ferries' buffers are filled partially to $\varepsilon \cdot b$, where $0 < \varepsilon \leq 1$. ε is the ratio of the time between ferry arrivals and the time to fill the buffer:

$$\varepsilon = \frac{\frac{2}{n}\left(\frac{l-2r}{v}+\frac{b}{R}\right)}{\frac{b}{R}} = \frac{2}{n}\left(1+\frac{l-2r}{v}\,\frac{R}{b}\right)$$

Packet Delay Calculation

In this section we calculate the average delay for one individual packet from source to destination analoguous to the chain-relay model.

(a)The First-on Case (worst-case delay) The delay this packet experiences comprises these time components.

i) *packet wait delay* until the next ferry arrives and stops at a distance r from the center of the source; this includes the actual travel time of the ferry as well as off-loading and loading at the destination, plus off-loading at the source. In the single flow case there is no loading at the destination or subsequent off-loading at the source. For n ferries this delay is on average as the n^{th} fraction of the cycle time less one buffer transfer time. This is not the average wait time since the ferry had just departed and a full cycle needs to be traveled.

$$t_{wait} = \frac{2}{n}\left(\frac{l-2r}{v}+\frac{b}{R}\right) - \frac{b}{R}$$

ii) *ferry loading time.* Even though the packet is loaded immediately into the ferry buffer, the packet still

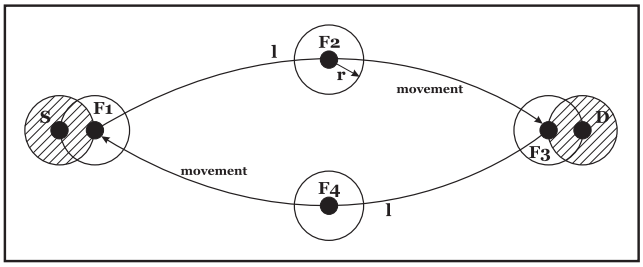

Figure 3: The Conveyor Belt Model

has to wait until all packets are loaded and the ferry is ready to depart.

$$t_{loading} = \frac{b}{R}$$

iii) *ferry transit delay* to destination. There will be no offloading delay since the packet is the first in the FIFO queue and can be handed to the application immediately.

$$t_{transit} = \frac{l - 2r}{v}$$

(b)The Last-on Case (best-case delay) A packet in the last-on case experiences the transit delay plus the offloading delay at the destination:

$$t_{best} = \frac{l - 2r}{v} + \frac{b}{R}$$

The average delay of packets from source to destination is calculated as the average of best case and worst case delays:

$$delay^{conveyor} = \left(1 + \frac{1}{n}\right)\frac{l - 2r}{v} + \left(\frac{1}{2} + \frac{1}{n}\right)\frac{b}{R} \quad (5)$$

2.4 Evaluation of proposed route designs

Each of our proposed route designs yields benefits over the other, depending on the environment it is deployed in. The following section compares both schemes in regards to throughput and packet delay. Environment variables include ferry speed v, ferry buffer size b, number of ferries n, data rate R, transmission radius r, and separation distance between sender and destination l.

Throughput

In the case of $n = 1$ ferries, both throughputs are equal, i.e.,

$$T = \frac{1}{\frac{2(l-2r)}{v\ b} + \frac{2}{R}}, \quad (6)$$

which validates the model analysis.

For $v \to \infty$ or $b \to \infty$ the chain model throughput becomes $\frac{R}{2}$ and the conveyor belt model yields $\frac{n \cdot R}{2}$.[3] Ferries with high buffer size can deliver more data per cycle in the conveyor belt model, since there are no intermediate buffer exchanges like in the chain-relay model. However, when using a fast radio, i.e., $R \to \infty$, the chain-relay model outperforms the conveyor-belt.

If communication range for each node is reduced, i.e., $r \to 0$, the conveyor belt model yields higher throughput. This suggests that when limited to short-range communication radios like Bluetooth, the chain-relay model should not be used.

The number of ferries such that $T = \frac{R}{2}$, which is the maximum data rate for the chain-relay model, is less for the conveyor belt model if $v > \frac{Rr}{b}$.

Delay

For $R \to \infty$ the chain model yields lower delay than the conveyor belt. When dealing with fast planes, i.e., $v \to \infty$, the conveyor belt model yields lower delay.

In summary, the conveyor belt model is superior with fast ferries, whereas the chain-relay model should be deployed with high data rate or long-range radios.

3 Algorithm Design

The previous section analyzed two models of possible ferry movement between one task node communicating to one data sink. Now we evaluate an algorithm which assigns ferry time to nodes in the case of multiple task nodes transmitting to one data sink.

3.1 Full Knowledge

The following scenario exemplifies the use of a single ferry to serve as communication enabler in a centralized, client-server star network. k stationary task nodes send traffic at their individual, but time-invariant, rates to a central server outside their coverage area. The ferry connects clients and server in a round-robin fashion, i.e., nodes are served one after the other. The ferry moves between client and server according to the conveyor-belt model. We assume full knowledge of the network parameters, including traffic flow rate f_i for each node i, distances d_i from node to server, maximum wireless data rate R at each node, buffer size b and constant speed v of the ferry. The ferry's radio can match any node's data rate, and the time for one buffer exchange is $\frac{b}{R}$. There is always a full buffer transmitted.

[3] $n_{opt} = 2$ in this case.

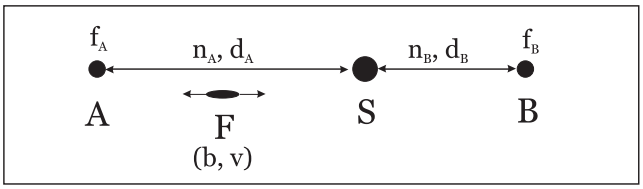

Figure 4: Problem notation: One ferry serving two nodes

The goal is to minimize the average packet delay in the network. The parameter to optimize for is the fraction of the total cycle time n_i that the ferry cycles between a given node i and the server before going on to serve the next node. A ferry can, for instance, spend more time serving nodes with higher flow rate requirement, than other nodes. Figure 4 shows the network setup with $k = 2$ nodes.

The objective function for packet delay in the network can be derived analogous to the conveyor-belt model by averaging over best case and worst case packet delay for each node. To consider all packets in the network, we further need to sum the delays over all nodes. Let each node's serving time, i.e., the total time it takes the ferry to cycle between node and server, be $n_i(2\frac{d_i}{v}+2\frac{b}{R}) = n_i\tau_i$. The total cycle time is $T_{cycle} = \sum_i n_i\tau_i$.

A packet experiences the shortest delay when it gets loaded onto the ferry last, i.e., $\frac{d_j}{v} + \frac{b}{R_j}$. If a packet arrives just after a ferry has departed, it incurs the longest delay, i.e., the cycle time to serve all other nodes, plus its own serving time. This delay is given by $3\frac{d_j}{v} + 2\frac{b}{R_j} + \sum_{i,i\neq j} n_i\tau_i$.

Averaging over both these delays yields the objective function:

$$d = \sum_{j=1}^{k} \left(2\frac{d_j}{v} + \frac{3}{2}\frac{b}{R_j} + \frac{1}{2} \sum_{i=1;\ i\neq j}^{k} n_i\tau_i \right) \tag{7}$$

The first two terms of (7), which do not depend on n_i, are positive and constant by definition, thus they would only add to the total delay. For simplicity, these terms can be neglected without affecting the solution, and (7) simplifies to

$$d = \frac{k-1}{2} T_{cycle}.$$

Note that this model weighs the delays equally by nodes. An alternative would weigh each node's delay by its traffic volume (i.e., its flow rate f_i). It turns out that this does not affect the results and so will not be used for simplicity.

To meet the traffic demand for each node, the following flow constraints need to hold.

$$f_i \cdot T_{cycle} \leq n_i \cdot b \ \forall\ i \in \{1, .., k\}, \tag{8}$$

$$n_i \geq 0. \tag{9}$$

Note that if the constraints are satisfied for any $\mathbf{n} = (n_1, n_2, \ldots, n_k)$ then they can be satisfied for any scaled version of $\mathbf{n}$. Similarly the delay can be minimized by scaling $\mathbf{n}$ so that $n_i \to 0$. Therefore for concreteness we let

$$\sum_{i=1}^{k} n_i\tau_i = T_{cycle} = 1. \tag{10}$$

Substituting (10) in (8) leads to

$$n_i \geq \frac{f_i}{b}. \tag{11}$$

Substituting (11) in (10) gives

$$\sum_{i=1}^{k} \left(\frac{f_i d_i}{v\ b} + \frac{f_i}{R_i} \right) \leq \frac{1}{2}. \tag{12}$$

Now even if $v \to \infty$, buffers need to be transmitted twice and (12) reduces to $\sum_i f_i \leq \frac{R}{2}$. If the data rate approaches ∞ the ferry still needs to fly back and forth twice to deliver data and (12) reduces to a constraint on flying time.

The fraction of time spent serving each node just depends on this node's flow requirement. The separation distance of node and server is not directly relevant. If only a few packets are transmitted from the node, the ferry visits this node less frequently, but instead delivers more packets on links with higher traffic flow.

Thus we have the following algorithm. First, the constraint in (12) is evaluated. If it is satisfied then the ferry can carry all flows successfully. Next the n_i are normalized to $n_i' = \frac{n_i}{\sum n_i}$ so that n_i' represents the fraction of time each node is visited. A simple stochastic algorithm would choose the next node to visit according to the probabilities $\{n_i'\}$. A deterministic algorithm would use a fractional visit counter c_i for each node to track visits.

1. Set $c_i = 0\ \forall\ i$.
2. While $\max\{c_i\} < 1$ set $c_i = c_i + n_i'\ \forall\ i$.
3. Let $j = \arg\max_i\{c_i\}$.
4. Visit node j and set $c_j = c_j - 1$.

5. When the ferry visit is finished, go to 2.

This specifies a fair algorithm in the sense that if all n_i are equal then each node will be visited once per cycle. If traffic was in the opposite direction from the central hub to the task nodes, then the scheduling would be trivial. The ferry would choose the task node with the most data waiting. This strategy is equivalent to the above deterministic algorithm.

3.2 Variable Data Rate Communication

All aforementioned approaches to determining the best location or route for a ferry were based on a simple communication model. In this model two nodes are able to communicate if the separation distance between them does not exceed a certain threshold. Once this distance threshold is exceeded, direct communication between the nodes is assumed impossible.

The model in reality is richer than this. Many digital wireless interfaces such as IEEE 802.11 have multiple data rates. When communicators are close the communication rate is high (e.g. 54 Mbps in 802.11g). When communicators are far the communication rate is lower (802.11g has 14 different rates between 54 and 1 Mbps). The rate decision is based on signal power to noise power ratio (SNR) and packet success measurements made over the channel. This phenomenon is an example of a general principle. The well known Hartley-Shannon Law states that the channel capacity C, which is the theoretical maximum rate at which information passes error free over a channel, is equal to

$$C = W \log_2(1 + \text{SNR}), \tag{13}$$

where W denotes the channel bandwidth. Depending on channel coding, actual data rates can be pushed very close to this theoretical maximum. The assumption of a wireless propagation model which links distance and SNR in the fashion $SNR = \frac{k}{d^e}$ leads to a relationship between data rate and distance

$$C = W \log_2(1 + \frac{k}{d^\epsilon}), \tag{14}$$

where k is a constant, and ϵ the path-loss exponent, usually between 2 and 4. The above equation turns the discrete data rate assumption in earlier works into a continuous function. Hereby, there exists an obvious trade-off between data rate and distance, which can be leveraged for providing improved connectivity of multiple nodes through ferry placement. The idea is that the ferry can exploit communication while moving between multiple nodes in order to more quickly transfer data.

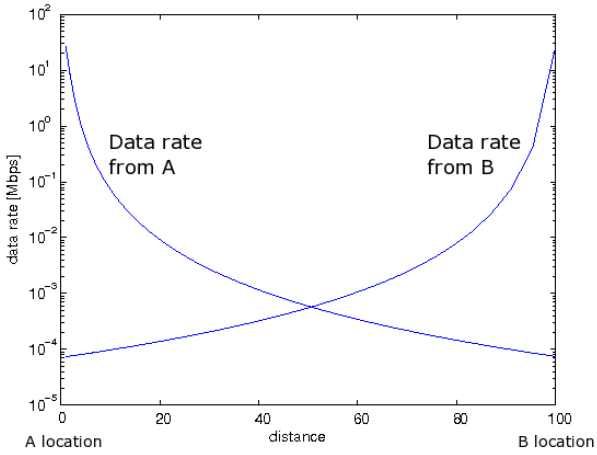

Figure 5: The possible data rates between ferry and nodes A and B as a function of ferry separation distance from A.

Figure 5 shows the achievable data rate between a ferry and node as a function of their separation distance for two nodes A and B.

In this paper we are concerned with the region when the ferries and nodes are relatively widely spaced. In this regime, SNR $\ll 1$ and we can approximate the Shannon capacity as:

$$R(d) = 1.44W\frac{k}{d^\epsilon} = R_0\frac{d_0^\epsilon}{d^\epsilon} \tag{15}$$

where R_0 is the bandwidth at a reference location d_0.

Scenario 1: One Single Ferry

Consider a node sending data to a sink. The nodes are separated by distance l. The optimal strategy for a single ferry is for the ferry to start at the midpoint, fly toward the node until its buffer is half full, then fly away from the node. Upon reaching the midpoint, its buffer will be full. Continuing to fly toward the sink, it will start to unload its buffer, until it has unloaded half of its buffer, then fly away from the sink. Upon reaching the midpoint, the rest of the buffer will have been unloaded and the transfer will be complete. In this way, the plane only moves as close to a node as is necessary to complete a transfer.

Consider two nodes separated by distance l. When a ferry is at distance y from a node and flying at velocity v, as it moves over a distance dy, it will relay $\frac{R(y)dy}{v}$ bits. So moving from the midpoint to within x of the node it will load:

$$B(x) = \int_x^{l/2} \frac{R_0 d_0^\epsilon}{v y^\epsilon} dy = \frac{R_0 d_0^\epsilon}{v(\epsilon-1)} \left(\frac{1}{x^{\epsilon-1}} - \frac{1}{(l/2)^{\epsilon-1}} \right)$$
$$= \frac{R_0 d_0^\epsilon}{v(\epsilon-1)x^{\epsilon-1}} \left(1 - \left(\frac{2x}{l} \right)^{\epsilon-1} \right)$$

Setting $B(x) = b/2$, setting $d_0 = l/2$ and solving for x yields;

$$x = \frac{l}{2} \left(\frac{1}{1 + \frac{b(\epsilon-1)v}{R_0 l}} \right)^{\frac{1}{\epsilon-1}}$$

The cycle time is then

$$T_{cycle} = 4(l/2 - x)/v$$
$$= \frac{2l}{v} \left(1 - \left(\frac{1}{1 + \frac{b(\epsilon-1)v}{R_0 l}} \right)^{\frac{1}{\epsilon-1}} \right)$$

The time $2l/v$ is the total time for the vehicle to fly from one node to the other and back. The quantity in the brackets is the fractional savings in time because of the transferred data enroute. To better see what is going on, we compute the throughput for the special case of free-space path loss, $\epsilon = 2$:

$$T = \frac{b}{T_{cycle}} = \frac{R_0}{2} + \frac{v \cdot b}{2l} \quad (16)$$

In this case we see that if v or b are large or R_0 or l are small then the throughput is one buffer per round-trip flight between nodes, $\frac{v \cdot b}{2l}$. If v or b are small or R_0 or l are large then the throughput is half the rate at the midpoint, $\frac{R_0}{2}$.

It would be instructive to compare this to the fixed radius communication model. In this model the data rate is R at radius r. Substituting this into the rate model yields $R_0 = Rr^\epsilon/(l/2)^\epsilon$ and:

$$T = \frac{2Rr^2}{l^2} + \frac{v \cdot b}{2l} \quad (17)$$

Let T_{vr} be the above throughput with variable rates, and T_{fr} be the fixed radius throughput in (6). We note that $T_{fr} \leq R/2$, while T_{vr} is unbounded as v or b increases. In the best case, when $l = 2r$, $T_{fr} = R/2$ which is less than $T_{vr} = R/2 + v \cdot b/(4r)$. While T_{vr} is generally superior, for large l they both have similar throughput of about $v \cdot b/(2l)$.

Scenario 2: Multiple Ferries

The superiority of the variable rate model is more pronounced in the case of multiple ferries. In the conveyor belt, more ferries can always be added by passing closer to the source and destination nodes where the rates are higher. In the chain relay, the effective l is shortened with each additional ferry and the communication is more into the regime where the variable rate method dominates. The conveyor belt model is trivial, so we will look at multiple ferries in the chain relay model.

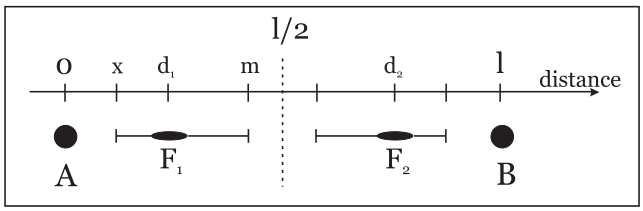

Figure 6: Multiple ferry scenario with $n = 2$ ferries.

We assume n ferries which serve evenly spaced division points between communication points A and B. Figure 6 shows the two ferry scenario. The ferry operation dictates that the gap between two adjacent ferries is being closed at twice the speed as the gap between a fixed endpoint and a ferry. However, throughput goes down. Looking at $n = 2$ ferries, we see that $d_2 - d_1 = 2d_1 = 2(l - d_2)$. Substituting $2l$ for l, $2v$ for v, and $R_0 = \frac{Rr^\epsilon}{(l/2)^\epsilon}$ in (16) this fact becomes obvious. Now, with n ferries the approximate throughput between adjacent ferries is given by substituting $v = 2v$ and $l = \frac{2l}{n}$ in (17), resulting in

$$T_{vr} = n^2 \frac{Rr^2}{2l^2} + n \frac{v \cdot b}{2l}.$$

So we see that throughput scales quadratic with the number of ferries serving the endpoints[4].

4 Related Work

Li and Rus [4] are one of the first to address deliberate trajectory changes to make message sending between two communicating nodes possible. Two algorithms are presented - one for full knowledge of node motions the other under partial knowledge. Each algorithm results in the local update of a node's trajectory. Errors in the estimation of neighbor node locations are studied, and the algorithms are simulated. Special attention is paid to the influence of the node's message relaying duty on the primary task of each node. They use an idealized propagation model, and have no consideration of multiple flows at a time.

[4]Recall this is the case where $\frac{l}{n}$ is large and (15) applies.

The first notion of message ferrying was developed by Zhao [7] who looks at routing of messages in partitioned wireless ad-hoc networks, where it is infeasible to route messages along pre-established routes. Instead, messages are *carried* by so-called "ferries" until they reach their respective destinations. Known traffic patterns are exploited to form efficient ferry routes that minimize delays and support bandwidth requirements. While at first looking at stationary ground nodes, in [8] the authors develop two algorithms for ferry route design in the mobile scenario. One ferry services all nodes. Nodes or the ferry can initiate the messaging scheme. Either the ferry location/trajectory is assumed to be known or the nodes are equipped with long range radios to notify the ferry of their locations. This way, the authors circumvent the problem of finding nodes and ferries. In [9] they extends the route design problem to include multiple ferries. Buffer constraints are not considered. Simulations are done in the network simulator ns-2. All path planning algorithms build upon the Traveling Salesman Approach.

Unlike our approach, the TSP solution will always find an optimal cycle path among all nodes. Thus every node is visited once per cycle. In the simple case of one outlying node with a low flowrate this will limit the network capacity and increase average packet delay in the network. The node will be visited in every cycle regardless of its location or flow requirement.

Shah et al. [6] deal with the problem of sensor data collection in sparse sensor networks. So-called MULEs pick up the data when in close range to the data-gathering nodes, buffer it for some time, and drop it off at wired access points. The main advantage is the power savings in the nodes, since only short-range transmission of data is necessary. An analytic model of the system is introduced, and scaling with regards to number of mules, sensors and access points is examined. Results are meant to provide guidelines for deployment of similar systems.

Merugu et al. [5] look at identifying recurring 'paths over time' in highly partitioned networks with explicit node movements. In these *store, carry, and forward* networks it sometimes is more efficient to carry data instead of forwarding it to the closest neighbor. Their work is based on an assumed periodicity of movement in most mobile nodes. The result is a space-time routing framework which relies on space-time routing tables where the next hop is selected from current as well as future neighbors. An algorithm for creating the routing tables as to minimize delay is introduced and empirically evaluated using simulations.

Chatzigiannakis et al. [3] introduce the notion of the *support* of the network, which refers to a subset of mobile nodes dedicated to message relaying. The movement of the support is controlled by a subprotocol. Two possible motion models are proposed: (a) the *snake* protocol, where support nodes stay pairwise adjacent at all times, and (b) the *runners* protocol, where each support node performs an independent random walk on a motion graph. It is shown that basic message passing can be improved especially in highly dynamic environments. The drawback of the approach is a limitation of the support movement along an abstract graph, which does not relate to real-world coordinates. Further, nodes in the runners protocol perform random movements which can not guarantee expedited delivery of high-priority messages.

5 Conclusion and Future Work

This paper analyzes two novel ways of designing ferry routes in a delay-tolerant network, namely the chain-relay model and the conveyor belt model. It is shown that each of the models has its merits, depending on the environment it is to be deployed in. A simple client-server star network, where one ferry serves multiple nodes according to the conveyor belt model, is studied and a delay-optimal scheduling algorithm is derived. Further, a new problem framework for data ferrying in delay-tolerant networks is described. It is argued that viewing the communication data rate as a continuous function of node separation distance can greatly enhance network performance. The design of mobility algorithms as well as transmission protocols within this framework will be the focus of future work.

References

[1] T. X. Brown, B. Argrow, C. Dixon, S. Doshi, R.-G. Thekkekunnel, and D. Henkel. Ad hoc uav-ground network (augnet). In *AIAA 3rd Unmanned Unlimited Technical Conference*, Chicago, IL, September 2004.

[2] T. X. Brown, S. Doshi, S. Jadhav, D. Henkel, and R.-G. Thekkekunnel. A full scale wireless ad hoc network test bed. In *Proceedings of ISART'05*, NTIA Special Publications SP-05-418, pages 51–60, March 2005.

[3] I. Chatzigiannakis, S. Nikoletseas, and P. Spirakis. Experimental evaluation of basic communication for ad-hoc mobile networks. In *Proc. of the 5th Annual Workshop on Algorithmic Engineering - WAE'01*, volume 2141 of *Springer Lecture Notes in Computer Science*, pages 159–171, 2001.

[4] Q. Li and D. Rus. Sending messages to mobile users in disconnected ad-hoc wireless networks. In *Mobile Computing and Networking (Mobicom)*, pages 44–55, 2000.

[5] S. Merugu, M. Ammar, and E. Zegura. Routing in space and time in networks with predictable mobility. Technical Report GIT-CC-04-07, Georgia Institute of Technology, March 2004.

[6] R. C. Shah, S. Roy, S. Jain, and W. Brunette. Data mules: Modelling and analysis of a three-tier architecture for sparse sensor networks. *Elsevier Ad Hoc Networks Journal*, (1):215–233, Sept. 2003.

[7] W. Zhao and M. Ammar. Message ferrying: Proactive routing in highly-partitioned wireless ad hoc networks. In *the 9th IEEE International Workshop on Future Trends of Distributed Computing Systems*, May 2003.

[8] W. Zhao, M. Ammar, and E. Zegura. A message ferrying approach for data delivery in sparse mobile ad hoc networks. In *MobiHoc'04*, May 24-26 2004.

[9] W. Zhao, M. Ammar, and E. Zegura. Controlling the mobility of multiple data transport ferries in a delay-tolerant network. In *Proc. of InfoCom 2005*, 2005.

Development of Performance Testing Methods for
Dynamic Frequency Selection (DFS) 5-GHz Wireless Access Systems (WAS)

Frank Sanders, Jeffery Wepman, and Steven Engelking
NTIA Institute for Telecommunication Sciences, Boulder, CO USA

Voice: +1-303-497-7600/3165/4430; fax: +1-303-497-3680
fsanders@its.bldrdoc.gov, jwepman@its.bldrdoc.gov, sengelking@its.bldrdoc.gov

Abstract. *Dynamic frequency selection (DFS) is an agile radio technology designed to allow wireless access systems (WAS) to operate in 5-GHz spectrum bands that are allocated on a primary basis to radiolocation systems (radars) without causing interference to radar operations. DFS is designed to accomplish this feat by detecting co-channel radar emissions and then avoiding or vacating any locally occupied radar frequencies. DFS technology thus promises to provide more radio spectrum for applications including multimedia transmission without denying use of that spectrum to existing users. Because the successful deployment of 5-GHz DFS technology in commercially available WAS devices depends critically on the ability of testing labs to verify that such products can detect co-channel radar emissions and vacate those channels, the development of adequate performance verification methods has been critical to the development of DFS technology as a whole. This paper summarizes the history of DFS spectrum allocation by the International Telecommunication Union (ITU), the specification of DFS performance parameters in a seminal ITU Recommendation, and the development in the United States of DFS performance verification test methods that can be applied to commercially produced DFS WAS products. That development has been carried out primarily by the U.S. Department of Commerce National Telecommunications and Information Administration Office of Spectrum Management and the NTIA Institute for Telecommunication Sciences, working in close coordination with other Federal agencies (including the Federal Communications Commission) and U.S. industry.*

1. Introduction

As a limited resource, radio spectrum should be used as efficiently as possible. To that end, a variety of schemes have been suggested in recent years to share spectrum between services that have historically been deemed to be incompatible. Two such services have been wireless communications (both fixed and mobile) and the radiolocation (radar) service. Radar receivers, being noise-limited in their performance, typically experience interference effects (especially manifested as lost targets) from high duty cycle signals (such as wireless communications) at interference-to-noise ratios in the receiver IF sections at levels ranging between -12 dB to -6 dB [1]. This effect is shown, for example, for a long-range, air-search radar in Figure 1, where targets that have a 90% probability of detection in the absence of interference begin to be lost at an I/N level of -10 dB. Prevention of such interference has been considered so challenging that these services have historically been allocated in separate spectrum bands to ensure that electromagnetic compatibility problems do not occur between them.

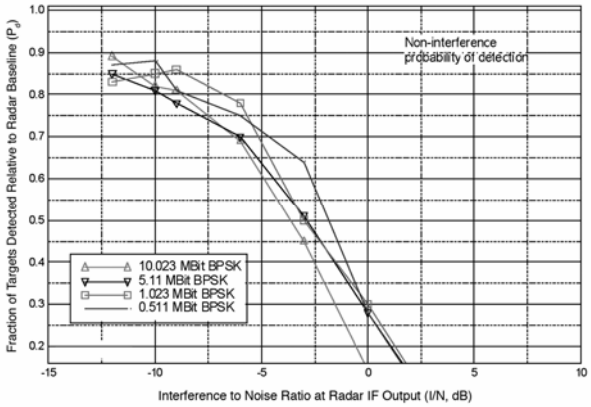

Figure 1. Example of the effects of interference on the probability of detection of targets by a long-range, air-search radar. Interference effects begin to occur at an I/N level of -10 dB.

This band-separation approach to spectrum allocation has been challenged somewhat in recent years. Another approach is now being attempted in the 5-GHz part of the spectrum, in which some operations of wireless communication devices are to share spectrum bands with radars. The primary solution is a technical approach called dynamic frequency selection (DFS). If

successful, DFS should allow wireless access systems (WAS)[1] to operate in radar bands through the detection and avoidance of frequencies that are used locally by radar transmitters.

This paper begins with the historic development of this new approach to spectrum sharing from a regulatory and technical standpoint. It then describes the development of techniques and protocols for testing and verifying the ability of commercially manufactured 5-GHz DFS WAS devices to detect radar emissions and avoid operations on those radar frequencies.

2. Example of the Use of Separate Spectrum Bands for WAS Versus Other Services

Separate spectrum allocations for different (and technically incompatible) services have often been successful. In the United States, for example, WAS that are compliant with the IEEE 802.11b and 802.11g standards are operated on an unlicensed basis in the 2400-2483.5 MHz band. This represents the lower part of the 2400-2500 MHz band (2.4-GHz band) that is allocated for industrial, scientific, and medical (ISM) uses; the other major occupier of the 2.4-GHz ISM spectrum is microwave oven emissions [2, 3].

The cut-off frequencies of 2400 MHz and 2483.5 MHz were empirically derived from measurements of microwave oven emissions at the Institute for Telecommunication Sciences (ITS)[2] as *limiting* frequencies between which microwave oven emissions might cause interference to reception of signals in the direct broadcasting satellite (sound) (BSS) service. Consequently, the BSS service was allocated on a primary basis to frequencies between 2310-2360 MHz and 2500-2690 MHz[3] [4, 5]. Just above the (former) upper BSS band, aeronautical search radars, such as are used at airports, have long been allocated on a primary basis between 2700-2900 MHz. The approach of designating separate band allocations for these three services (ie., lower BSS; ISM-WAS-microwave ovens; upper BSS; and aeronautical radars), with so-called guard band separations between them of 40 MHz, 16.5 MHz, and 10 MHz, respectively, have worked to ensure electromagnetic compatibility between their operations.

3. The Development of the DFS Concept for Sharing Spectrum Between the Radar Service and WAS

In the 5-GHz part of the spectrum, regulatory requirements and band allocations have varied widely around the world for WAS 802.11-compatible devices. This variation has tended to make it difficult to manufacture IEEE 802.11a-compatible equipment for the 5-GHz bands [6]. Moreover, radar services for the purposes of radiolocation and radionavigation are allocated in the US on a primary basis in a set of bands running contiguously between 5250-5850 MHz. The use of this band by radars is substantial, as documented in an NTIA Report [7]. But unlike the example of the 2.4-GHz band described above, which does not overlap a radar band, the 5725-5875 MHz ISM band partly overlaps radar band allocations.

Internationally, the 5-GHz radar band allocations are somewhat more complicated, but they likewise overlap the 5-GHz spectrum that is perceived as desirable for use by WAS. These internationally recognized 5-GHz primary allocations for radar bands are as follows [8]: 5250-5350 MHz is allocated to the radiolocation service on a primary basis; 5470-5650 MHz is allocated to the maritime radionavigation service on a primary basis; ground-based meteorological radars operate between 5600-5650 MHz on a basis of equality with stations in the maritime radionavigation service; and 5650-5725 MHz is allocated to the radiolocation service on a primary basis.

It is desirable to make 802.11a WAS devices available on a worldwide basis, if possible, in 5-GHz spectrum without an undue burden on manufacturers to produce many different varieties of 5-GHz WAS to meet varying 5-GHz spectrum requirements from one governmental administration to the next. (Such harmonization would also make it easier for international travelers to use 5-GHz WAS devices as they move across boundaries between administrations.) The overlap of primary radar spectrum allocations with the 5-GHz ISM band has made it desirable to find technological methods to allow 5-GHz WAS (which have historically tended to use ISM spectrum for their operations) to share spectrum with radars. An ancillary goal has been to ensure a spread of WAS loading across the 5-GHz spectrum, in order to reduce the aggregate WAS emission levels at the frequencies of the satellites of the fixed satellite service and earth exploration satellite service (EESS active).

---

[1] Usually more specifically described as wireless local area networks (WLANs), radio local area networks (RLANs), or Wi-Fi systems.

[2] ITS, located in Boulder, CO, is part of the U.S. Department of Commerce, National Telecommunications and Information Administration (NTIA).

[3] The 2500-2690 MHz allocation for BSS was recently deleted by the FCC, but is mentioned here because it illustrates the principle of separation of services between bands.

The needs for 5-GHz spectrum harmonization and development of technical methods to allow sharing of 5-GHz spectrum between radars and WAS were addressed by the International Telecommunication Union Radiocommunication Sector (ITU-R) at the World Radio Conference of June-July 2003 (WRC-03). At that conference, the ITU administrations considered the dual needs to harmonize frequencies in the bands 5150-5350 MHz and 5470-5725 MHz for the mobile service to facilitate the introduction of wireless access systems (WAS), while still protecting the radars that operate in the bands 5250-5350 and 5470-5725 MHz. The document that was produced on this subject at WRC-03 was ITU-R Recommendation M.1652 [8]. It recommends technical mitigation techniques for government administrations to use in facilitating spectrum sharing between radars and WAS devices in the referenced 5-GHz radar bands; DFS is the technique that is primarily recommended and described in M.1652 as the solution to the problem.

4. DFS as Described in ITU-R M.1652

ITU-R M.1652 was developed through lengthy and difficult technical work by experts from many administrations in the ITU-R over the course of several years, notably in the Joint Task Group 8A/9B (JTG 8A/9B) but also to some extent in Working Party 8B (WP 8B). Thus, the content of ITU-R M.1652 actually represents the culmination of work that began many years before 2003 in various ITU-R working groups.

The Recommendation states that harmonized frequencies in the bands 5150-5350 MHz and 5470-5725 MHz for the mobile service would facilitate the introduction of WAS, including RLANs, and goes on to call for protection of radar operations in the bands 5250-5350 and 5470-5725 MHz through the implementation of technical methods for mitigation of interference effects in radar receivers. It contains detailed technical annexes that describe the characteristics of 5-GHz radars (as submitted over a number of years from several ITU-R member administrations including the US). The DFS method for protection of these radars from interference by WAS occupies the bulk of the remaining document annexes.

The goal of DFS, as described in M.1652, is for WAS-type communication systems to automatically sense the presence of radars on locally occupied frequencies and to vacate those frequencies in a timely fashion. A fundamental assumption behind DFS is that most 5-GHz radars scan their beams through space in some fashion, and that such radar emissions will be most detectable at the moments in which they scan their transmitted beams across WAS locations.

The criteria for detection of radar emissions by a DFS-equipped WAS are described in M.1652 as follows: at the location of the WAS, the minimum detection threshold for radar signals is set at -62 dBm for WAS devices with a maximum effective isotropic radiated power (EIRP) of less than 200 mW and -64 dBm for devices with a maximum EIRP of 200 mW to 1 W averaged over 1 μs.

The detection thresholds are to apply to power levels within the WAS receiver circuitry that are normalized to the output of a receiver antenna with a gain of 0 dBi, in the bandwidth of the WAS device. For example, suppose that a WAS device operates at an emission level of 100 mW, with an antenna gain of 2.5 dBi and internal circuit gain of 6 dB between the antenna and the detector circuitry. Then a radar signal will be registered in the detector circuitry at a level that is (2.5+6)=8.5 dB higher than if it were monitored with an isotropic antenna through a receiver with zero gain. Consequently, the internal detection threshold setting for such a device would not be -62 dBm, but rather 8.5 dB higher, at -53.5 dBm.

Recommendation M.1652 only calls for the radar detection functionality (hardware and software) to be built into the WAS controller, but in such a way that *every* transmitter in the WAS will be guaranteed to cease transmissions when the controller is alerted to the presence of a radar signal. In practice, for example, a WAS consisting of an access point (AP) and multiple client units might only perform radar detection through a monitor incorporated into the AP and its computer. When a radar emission is detected by the AP, that unit would in turn alert all of the clients to stop transmitting on the existing frequency and move rapidly to another frequency.

M.1652 recommends that the detection of radar signals should occur under two different conditions: both when the WAS is initially activated, and also subsequently while the WAS is operating. When the WAS is initially turned on, it is supposed to monitor a selected radio channel for radar emissions for 60 seconds. If this channel availability check detects any radar signal according to the predetermined criteria (as given above), then the WAS controller must not allow the system to operate on that channel for at least the next 30 minutes, and instead must attempt to begin operations on a different channel (necessitating yet another channel availability check on the new channel, of course).

The channel availability check lasts for 60 seconds because some radars, especially some meteorological radars, only make a complete 360-degree scan about once per minute, and the availability check must last at least long enough to be able to intercept such emissions. Most other types of 5-GHz radars perform 360-degree beam scans in less than 60 seconds, and will be detectable by a 60-second interval for channel availability monitoring.

As an additional protection for some radars, any channel between 5600-5650 MHz that has been previously flagged as containing a radar signal may not be used without 10 minutes of additional, continuous monitoring. (Some administrations, including the US government, simply may not allow any operations in that band at all.)

After the WAS begins to operate on a given frequency, the system must continue to monitor for radar signals while it is transmitting and receiving traffic. This so-called in-service monitoring for radar signals is supposed to be performed during quiet intervals between data bursts. It is crucial for two reasons. First, many 5-GHz radars are mobile, and may move into an area, using a WAS channel, after a WAS has already begun to operate. Second, even fixed 5-GHz radars may not operate continuously; such a radar could be inactive when the WAS is turned on, but could then become active on the WAS channel at some point in time after the WAS has begun operations.

When a channel must be vacated due to detection of a radar signal, M.1652 recommends that all transmissions on that channel should cease within 10 seconds. But normal traffic on the channel is supposed to cease within 200 ms, and not more than a total 100 ms of that interval should actually be occupied by data traffic. Subsequently, the remainder of the 10-second shut-down interval is supposed to be occupied by not more than a total of 20 ms of intermittent communications devoted to managing the vacating of the channel (and presumably the coordination of the transfer of data operations to another radio channel) by all of the units in the WAS. The WAS message announcing the need to cease transmissions on the radar-occupied channel is supposed to be broadcast repeatedly to ensure that the channel in question is totally vacated by all units.

Subsequent to detection of a radar signal on a WAS channel, M.1652 requires the system to vacate that frequency for at least 30 minutes (called the non-occupancy period) before attempting any more operations on that channel. Instead, the system must identify and use another channel. But M.1652 states that no frequency should be utilized by the WAS until the 60-second availability check has been performed. To fulfill this recommendation while continuing its transmissions in a seamless manner, a WAS would have to either jump to a non-DFS channel (that is, one that is not in a radar band) upon radar detection, or else would have to always monitor at least one back-up channel for at least 60 seconds at a time while simultaneously transmitting traffic on the active WAS data channel.

5. Tradeoffs in DFS Designs

From the description of DFS functionality in M.1652, the question naturally arises as to whether a power detection threshold alone is adequate (in practice) to discriminate between radar emissions versus other energy on a channel, especially during in-service monitoring. Given that some noise may occur due to non-radar emissions, and that unwanted 5-GHz emissions might even originate from a neighboring (but uncoordinated) WAS, it is likely that WAS manufacturers may use more parameters than mere power detection to determine whether or not radar emissions are occurring on their systems' channels. For example, some sort of pattern detection algorithm might be invoked to search for particular intervals between a non-WAS series of incident pulses during in-service monitoring. This would help to avoid false alarms that could be caused by triggering on random noise pulses. But that approach would be complicated by at least three facts: radar pulse repetition intervals (PRI) vary widely from one radar model (and operational mode) to another; administrations will not usually release details of radar emission parameters such as PRI for particular radars; and in any event the intervals between radar pulses are not always fixed, but rather may vary (that is, may be jittered or staggered) in time.

Another possible method for verifying that radar pulses are present and avoiding false alarms might be to discriminate on the basis of radar pulse width versus WAS data packet lengths. But this solution would be complicated by problems similar to those of discriminating on PRI: pulse widths emitted by radars will vary widely depending upon the model type and operational mode; administrations will not usually release detailed information about radar pulse widths; and finally, some radar pulse widths are comparable in length to some WAS data packets.

In reality, it is likely that WAS DFS algorithms in deployed systems may typically use some sort of composite information about received amplitudes, detected pulse widths, intervals between incident pulses, and the number of pulses observed sequentially in some predetermined interval of time to adequately

discriminate between WAS traffic and radar signals. Whatever method is used, the key to designing a successful DFS algorithm will be to implement a methodology that strikes a balance between avoiding the generation of an excessive number of false alarms (which would then cause an unnecessarily large number of channel hops by the WAS), while at the same time being sufficiently sensitive to genuine radar pulses as to pass product testing requirements.

It should be emphasized that 5-GHz radars operated by various administrations vary greatly in their operational characteristics, and that administrations will not ordinarily make the details of these characteristics publicly available. Even if they did, the sheer variety of PRIs, pulse widths, beam scanning characteristics, frequency hopping capabilities, and pulse coding features that can occur in 5-GHz radars would make such an accounting nearly useless for purposes of detecting radar signals. Moreover, the future development of new radar modes in existing systems, plus the development of entirely new radar systems, would tend to render moot the desirability of obtaining details of radar emissions for the purpose of designing DFS radar-detection algorithms.[4]

Whatever sensing method is used, it is understood at present that manufacturers of WAS devices that will be available in the US will not be requested by NTIA to disclose the algorithms by which they detect the presence of radar signals. Instead, they will only be required to successfully pass performance testing with specified types of pulses at the incident power levels of -64 dBm or -62 dBm, in accordance with the ITU requirements described above.

Another inherent tradeoff in DFS algorithms is between the amount of time occupied by data traffic versus the amount of time that can be devoted to monitoring for radar signals. Radar signals can only be observed when quiet intervals occur between WAS data packets. The higher the traffic level, the less time there is to monitor for radar pulses. To understand the problem of detecting a burst of radar pulses as a radar beam sweeps across a WAS location, consider the following parameters that would be representative of a typical radar emission: the pulse width is 1 μs; the pulse repetition interval is 1 ms, the beamwidth lasts about 20 ms between the 3-dB points; and the beam-scanning interval has a periodicity of 5 sec. To obtain a 100 percent probability of detecting such a signal on the first scan of the radar beam across the WAS location, the detection algorithm would have to sample for radar pulses at least once every 20 ms (so as to never miss the radar beam-scan event that will inevitably occur once every 5 sec), and each of those monitoring events would have to last at least 3 ms (so as to extend through the interval of several radar pulses). Monitoring for at least 3 ms every 20 ms, the monitoring duty cycle would be (3/20) = 15 percent; this percentage of time would be unavailable for data traffic. As significant as this duty cycle would be, the need to perform such monitoring at least every 20 ms would represent a severe restriction on the maximum allowable length of data packets (which would of course never be able to exceed 20 ms in this example).

This example serves to demonstrate that the probability of detecting radar pulses during any given radar scan across a WAS location will realistically be less than 100 percent per radar beam-scan event; the actual probability of radar detection (on a per-scan-of-the-radar-beam basis) will be proportional to the ratio of the duration and frequency of the monitoring periods to the duration of radar beamwidth (such beamwidths typically being on the order of 10 ms to perhaps 100 ms) and the frequency of the radar beam-scanning behavior. Indeed, it is partly because this in-service monitoring detection probability cannot be expected to achieve 100 percent that Recommendation M.1652 specifies that DFS-equipped WAS devices should monitor *continuously* on the selected frequency for at least one full minute on start-up, prior to the beginning of the first data transmission.

6. Testing the Performance of DFS-Equipped WAS Devices

There are inherent uncertainties in understanding the performance of DFS-equipped WAS devices on a purely analytical basis, especially since the algorithms used by particular systems do not have to be disclosed by manufacturers and thus are not available for analysis. Therefore, a critical imperative for the approval of these devices has been to develop a method for testing and verifying their ability to sense (and respond to) radar signals on their operational frequencies.

This work is being moved forward in three stages. The first stage has been to develop and demonstrate a laboratory test bed for verifying the performance of prototype DFS-equipped WAS devices; the second has been to demonstrate the performance of some prototype devices under actual field conditions near a deployed 5-GHz radar; and the third will be to set up permanent testing systems in laboratories where actual certification

[4] In the US, rules have been developed to prevent any end user from accessing the WAS device algorithms or extracting *any* information about detected radar signals.

and acceptance tests will be performed on a long-term basis. The first two stages of work have been accomplished by groups within NTIA's OSM and ITS organizations, working in close coordination with the Federal Communication Commission (FCC), other Federal agencies, and industry personnel. The third stage of this effort will be implemented by the FCC.

It was determined about three years ago that the best approach for testing would be to expose DFS-equipped WAS devices to a variety of pulses of electromagnetic energy that would be representative of 5-GHz radar signals, consistent with the pulse width and pulse repetition parameters that are described in one of the Annexes of Recommendation M.1652.[5] It was decided that the parameters of those testing pulses would be divulged to manufacturers prior to testing, but it was further specified that under no circumstances were manufacturers to pre-program their prototype devices to search specifically for those pulse parameters. Table 1 contains pulse parameters that have been used for some of the testing at the ITS laboratory.

Table 1. Parameters of some of the radar pulses used for DFS laboratory testing in the US.

Radar test signal	Pulse repetition frequency (pps)	Pulse width (μs)	Burst length (ms)/no. of pulses	Burst period (sec)	Frequency hopping rate
Fixed frequency radar signal 1	700	1	26/18	10	n/a
Fixed frequency radar signal 2	1800	1	5/10	2	n/a
Frequency hopping radar signal	3000	1	100/300	10	1 kHz

During 2003, NTIA personnel at OSM and ITS developed a testing system that would be used to transmit these test radar pulses at representative, prototype DFS devices at frequencies near 5-GHz. As a result of considerable discussion about the desirability of coupling the test radar pulses into the DFS receivers via hardline connections versus through radiative coupling, ultimately a decision was made to perform radiatively coupled testing. This decision resulted partly from a desire to test the DFS devices in as close to a true operating mode as possible, and partly because of the practical consideration that WAS devices do not generally incorporate removable antennas, and thus it is not ordinarily possible to disconnect the antennas of such devices (as would be necessary to accomplish hardline-coupled testing) without inflicting some sort of damage to the devices.

As finally completed, the ITS 'radar' transmitter system is built around a vector signal generator (VSG) and a fast-switching programmable microwave synthesizer that are programmed with sets of pulse widths, pulse repetition intervals, operating frequencies, and pseudo-random frequency-hopping parameters that had been previously specified and agreed through inter-agency and industry coordination.[6] The system is capable of generating any of the specified radar pulses at RF frequencies near 5-GHz. The system has been operated inside a laboratory using a broadband horn antenna that is vertically polarized. At a distance of 3 m, a platform is set up for the DFS devices that are to be tested.

At the beginning of a test series, a calibrated, vertically polarized horn antenna is placed on the receiver platform in lieu of a DFS device, and the horn is connected to a spectrum analyzer. The spectrum analyzer is adjusted to 1-MHz Gaussian bandwidth (specified at the 3-dB points) and the analyzer's detection mode is set to positive peak. For the purposes of prototype-demonstration testing, the output power of the 'radar' is adjusted to produce a received power level somewhere between about -58 dBm to -60 dBm in the spectrum analyzer, as adjusted for the gain of the horn (typically about 10 dB relative to isotropic) and the loss in the cable between the horn and the analyzer.[7]

When the incident radar power level has been calibrated at the receiver platform, a prototype DFS-equipped AP unit is placed on the platform and is connected to its controller computer. One WAS client (with its own computer) is placed at a remote location in the laboratory, and data traffic is initiated between the AP and the client. (Ad hoc (client-to-client) networks have not been tested to date.) The traffic has been public-domain MPEG-2 video and wav-type audio files. The

[5] It is emphasized, however, that none of the sets of pulse parameter characteristics used in DFS testing in the US have replicated the parameters of any particular 5-GHz radars, nor will they.

[6] The radar waveform parameters are contained in the FCC 5-GHz Report and Order (FCC docket 03-122 at http://gullfoss2.fcc.gov/prod/ecfs/comsrch_v2.cgi).

[7] This power value is of course slightly higher than the -62 dBm to -64 dBm levels specified in M.1652. This is because the NTIA work has been intended to demonstrate the feasibility of the testing hardware and software; final acceptance tests performed by the FCC will presumably use power levels that will match the specified detection thresholds.

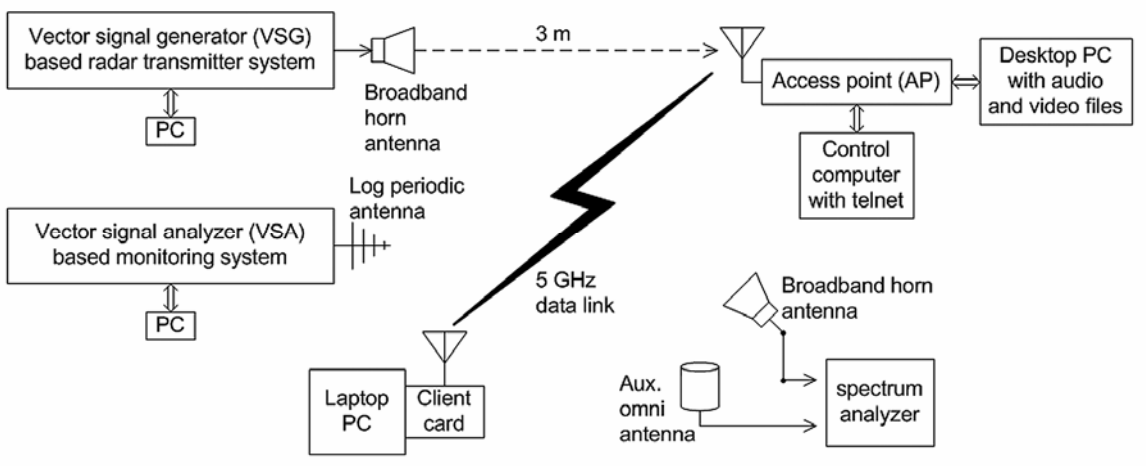

Figure 2. Overall setup of the DFS testing system that has been developed and used by NTIA. The VSA and spectrum analyzer monitoring antennas are located off-axis from the radar pulses and the 5-GHz data link.

video files have achieved an average of roughly 50 percent channel loading, and the wav files somewhat less; the video files have been used in most tests because they produce the highest loading. The overall test setup is shown in Figure 2.

Both the start-up and the in-service monitoring modes of the DFS device are tested with this set-up. In the start-up mode, a monitoring antenna connected to a spectrum analyzer is used to first of all verify that the WAS does in fact remain quiet for the first minute of its operation, with data traffic commencing at the conclusion of the 1-minute interval.

Subsequently, the AP-client start-up is repeated, but with a single burst of about 15 radar pulses transmitted at the calibrated power level once during the 1-minute start-up monitoring period. In these tests, the radar burst is either transmitted 6 sec after initial start-up or else 6 sec before the end of the 1-minute interval. For these tests, the spectrum analyzer is used to verify that the DFS does *not* initiate any operations on that frequency at the end of the 1-minute start-up interval.

With start-up behavior verified, the tests move on to verification of DFS performance during in-service monitoring. For that testing, the 1-minute start-up monitoring is suspended to speed the testing rate.

For in-service monitoring tests, the AP-client traffic is established and then a burst of about 15 radar pulses is fired on the WAS frequency if the radar mode is non-frequency hopping. (In pseudo-random frequency-hopping radar modes, a much longer series of pulses must be transmitted. This is necessary because a long interval may elapse before any of the radar pulses occur on the WAS frequency.) Every time a set of radar pulses is transmitted, three actions are taken: First, it is noted whether the AP did or did not react to the radar burst. Second, a spectrum analyzer connected to a monitoring antenna and operating in a zero-hertz span mode (that is, operating in the time domain) is used to provide a coarse verification of both the presence of the radar pulses and the shut-down (if any) of the AP-client traffic on the radar frequency. Third, a vector signal analyzer (VSA) connected to a second monitoring antenna is used to record the shut-down behavior of the AP-client traffic in enough detail to meet the specifications of Recommendation M.1652.

Furthermore, at least one in-service monitoring test is performed in which the WAS is exposed to a burst of radar pulses, and in which the WAS RF output is then monitored for the next 30 minutes to ensure that it does not attempt to use that frequency again during this so-called non-occupancy interval.

The spectrum analyzer provides low-resolution (millisecond time resolution) data traces virtually instantaneously with each radar burst, over a time interval that can last a few seconds or longer. In contrast, the VSA can provide data with microsecond time resolution (which is required to verify DFS conformance with testing requirements), but needs many minutes to provide these data, and cannot take such data for such a long period as the spectrum analyzer. Because of the long time required to generate its data outputs, the VSA is not used to record shut-

down behavior of the WAS for every radar burst; instead, the VSA is used to record representative samples for a few radar bursts in each radar mode. The spectrum analyzer, in contrast, is used to verify DFS shut-down behavior (or lack thereof) for every radar burst. Together, the combination of the spectrum analyzer and the VSA have been found to provide an adequate set of DFS test-performance verification data. Dedicated computers are used to control each of these instruments and to record information from them during testing.

A statistically significant number of radar bursts are transmitted in each radar mode, ultimately providing the probability of detection of radar emissions by each prototype DFS device. It has been observed that DFS-equipped WAS devices usually detect single-frequency radar signals more successfully than frequency-hopped radar signals.

It is emphasized that the tests performed with this system in the spring of 2004 and the late summer of 2005 have not been performed to certify or accept any particular DFS devices for sale in the US. Rather, their purpose was to demonstrate the feasibility and utility of this system for such testing. All but one of the tested devices have been based on 802.11 architecture; the remaining system has used a frame-based architecture in which the talk/listen ratio is user-controlled.

A testing system identical to the one developed at ITS might be used for actual acceptance tests, but other parties have been invited to propose (and demonstrate) the utility and feasibility of alternative approaches.[8] The system described here is not the only possible approach, but to date it is the only system that has been publicly demonstrated and reported.

In the second testing stage, some prototype DFS devices were tested in proximity to an operational 5-GHz radar at a field location. For these tests, the monitoring and verification equipment were again a combination of spectrum analyzer and VSA. But for these tests, the manufacturers were *not* informed of the radar pulse parameters that were transmitted by the radar.

In addition, the performance of that radar has been tested in the presence of an aggregate emission from DFS-equipped 5-GHz WAS devices. For those tests, aggregate WAS emissions were recorded with a VSA at ITS. At the field location, those recordings were radiated from a VSG through an antenna and into the radar receiver; the effects were noted as a function of the power level of the aggregate DFS emissions.

7. Preliminary DFS Testing Results

Figure 3, from the first round of testing in Boulder in 2004[9], contains time-domain data that show a burst of radar pulses and the cessation of operation of a DFS-equipped WAS before the burst has finished. This figure is typical of the spectrum analyzer data obtained in those tests. The VSA data show qualitatively similar information at much higher time resolution.

The results of the initial bench tests in 2004 showed that the 5-GHz devices failed to achieve a good rate of radar signal detection. Overall, between all the manufacturers the radar detection capabilities of the devices tested were moderate at best and the radar detection was highly dependent upon the amount of loading of the data channel. That is, detection occurred at a higher rate when the audio file was being streamed than when the video file was being used.

A key finding of the first round of testing was that the devices were not able to detect radar pulses that were comparable in length to typical 802.11 data packets. The devices apparently tended to misidentify long radar pulses as corrupted 802.11 data packets.

The characteristics of radars that use longer pulse widths are contained in ITU-R M.1652 and these radars must be protected and detected in a timely manner. It is understood by the authors that WAS-against-radar DFS tests performed by other government administrations have drawn similar results and conclusions. Those tests have used radar parameters similar to those that were used in the NTIA bench tests (personal communications[10]).

The second round of DFS testing in the US in 2005 indicated that significant improvements had been achieved in the performance of the DFS radar-sensing capabilities against the types of radar parameters listed in ITU-R M.1652, although frequency-hopping radar signals, which were difficult to detect in the first round of tests, were still difficult to identify in the second round.

[8] Indeed, it is possible that DFS acceptance testing may ultimately be performed with hardline-coupled radar pulses rather than radiatively-coupled pulses.

[9] This first round of testing at ITS was done in two phases, in January and May 2004.

[10] ITU-R Working Party 8B meetings, Geneva, in March and September 2005.

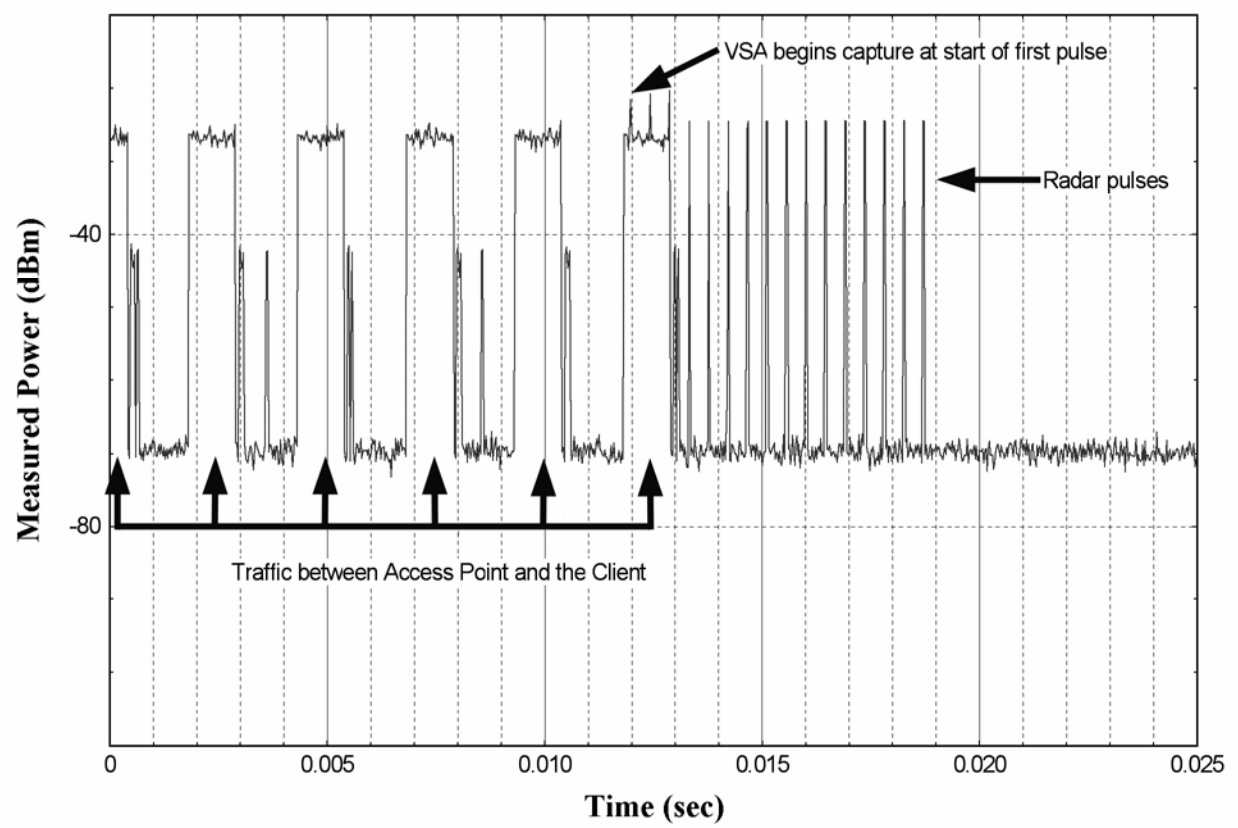

Figure 3. Time domain data from DFS testing, showing the cessation of data traffic on a DFS-equipped WAS channel when a burst of radar pulses occurred.

8. Next Steps for Implementation of DFS in the US

With the first and second stages of DFS testing (that is, laboratory demonstrations with replicated radar pulses and field tests against an operational radar) having been completed on prototype DFS-equipped WAS devices, the next steps in the US will be for the FCC to complete the rules for DFS performance criteria and testing protocols, and to implement systems for actual compliance testing. NTIA is providing all results of its work to the FCC to assist in this process.

9. References

[1] F.H. Sanders, "Technical challenges to spectrum sharing between radar and non-radar (communication) systems," In: "Proceedings of the International Symposium on Advanced Radio Technologies, March 1-3, 2005," NTIA Special Publication SP-05-418, Mar. 2005, pp. 21-29.

[2] P.E. Gawthrop, F.H. Sanders, K.B. Nebbia, and J.J. Sell, "Radio spectrum measurements of individual microwave ovens," NTIA Report in two volumes: 94-303-1 and 94-303-2, Mar. 1994.

[3] M.G. Biggs, F.H. Sanders, B.J. Ramsey, "Measurements to characterize aggregate signal emissions in the 2400-2500 MHz frequency range," NTIA Report 95-323, Aug. 1995.

[4] C.A. Filippi, R.L. Hinkle, K.B. Nebbia, B.J. Ramsey, and F.H. Sanders, "Accommodation of broadcast satellite (sound) and mobile satellite services in the 2300-2450 MHz band," NTIA Technical Memorandum TM-92-154, Jan. 1992.

[5] *Manual of regulations and procedures for federal radio frequency management*, National Telecommuni-

cations and Information Administration, U.S. Department of Commerce, May 2003 edition, revised Sep. 2005.

[6] J. Stancavage, "Wi-Fi popularity extends to 5-GHz," *Wireless Systems Design*, Penton Media Inc., Apr. 2004. Available electronically at URL: http://www.wsdmag.com/Articles/ArticleID/7915/7915.html.

[7] F.H. Sanders, "Measured occupancy of 5850-5925 MHz and adjacent 5-GHz spectrum in the United States," NTIA Report 00-373, Dec. 1999.

[8] International Telecommunication Union Radiocommunication Sector, "Dynamic frequency selection (DFS) in wireless access systems including radio local area networks for the purpose of protecting the radiodetermination service in the 5 GHz band," ITU-R Recommendation M.1652, June 2003.

Acknowledgments

The authors wish to thank Robert Sole of NTIA's Office of Spectrum Management for his important participation in and support of this work.

Impulse Radio Transmitter using Time Hopping and Direct Sequence Spread Spectrum Codes for UWB Communications

Antonio Mollfulleda
Centre Tecnològic de
Telecomunicacions de Catalunya (CTTC)
Tel.0034 93 6452920
Fax 0034 93 6452901
antonio.mollfulleda@cttc.es

Joan A. Leyva
Centre Tecnològic de
Telecomunicacions de Catalunya (CTTC)
Tel.0034 93 6452917
Fax 0034 93 6452901
joananton.leyva@cttc.es

Lluís Berenguer
ITOWA, Investigation Total Ware
Tel. 0034 93 6452917
Fax 0034 93 6452901
lluis.berenguer@cttc.es

Abstract— This work presents an UWB transmitter for reduced data rates. Time Hopping (TH) and Direct Sequence (DS) spread spectrum codes are used to improve the spectral efficiency below the FCC mask. The pulse generator is composed of a Class-S digital pulse amplifier and two Step Recover Diode pulse generator. An Inverter/Non-Inverter is introduced to implement the DS-SS code. In the last stage, a Wilkinson Power Combiner is presented to feed a single antenna. Theoretical and measurement results show that the PSD reaches a maximum of -45dBm/MHz and present a peak-to-average ratio of about 10dB.

Index Terms—Ultra-Wideband, Step Recovery diode, Class-S amplifier, Inverter / Non-Inverter

I. INTRODUCTION

The activities on Impulse Radio (IR) communication systems started on 1973 with the Ross Patent. In this system the carrier of information signal is a pulse train. Since then, communication systems based on IR have been used on radar, imaging and particularly on military systems as a low probability detection system. One characteristic of all this systems is that they are based on short duration pulses, associating IR system to wideband or Ultra-Wideband (UWB) systems. In 1998 the Federal Communications Commission (FCC) defined Ultra-Wideband systems as any communication based on a signal with a fractional bandwidth greater than 20% of the carrier frequency or with a bandwidth greater than 500 MHz. This definition includes both, carrier-based systems as well as impulse radio systems. In 2002 the FCC published a regulatory report allowing UWB signal transmissions with radiated Power Spectral Density (PSD) below –42dBm/MHz in the 3.1GHz-10.6 GHz frequency range. Following this report, a number of emerging commercial applications have revived significant research and development. Considering UWB systems there are two different possibilities: high data rate for short-range distances, as Bluetooth alternative; or, low data rate for high-range distances (up to 300 meters). Second chance is the one selected for the presented transmitter. In any case techniques for generating narrow pulses and UWB signals presenting efficient usage of the available power below the FCC mask is a hot research topic.

The transmission quality of any communication system depends on the received Signal to Noise Ratio (SNR), which in FCC compliant UWB systems, is strongly related to the efficient usage of the spectrum. However, the PSD of a generic IR system presents high power spectral lines that force to reduce the transmitted power in order to keep the radiated signal well bellow the mask limitations. The spectral lines are a consequence of cyclo-stationary property [1]. Using spreading codes in which more than one pulse per symbol is transmitted the level spectral lines can be reduced. Two types of codes can be used in traditional IR systems. In Time-Hopping Spread Spectrum (TH-SS) codes the pulse position is randomized inside some time interval known as a time hopping interval. The main purpose of the TH codes is to reduce the effect of the cyclo-stationary property. In Direct Sequence Spread Spectrum (DS-SS) codes the pulse sign is randomized. The DS codes are useful because from a statistical point of view it cancels the expected value of the pulse amplitude.

In this paper we present the architecture of an UWB transmitter. In section II IR signal is analyzed. Following, the design of an IR transmitter using TH-SS and DS-SS codes to reduce the effects of the spectral lines is presented in section II. Finally, the measurement results achieved are shown and the conclusion of this work is exposed.

This work is partially supported by the Spanish Ministry of Industry Tourism and Trade under contracts FIT-330200-2004-281 (PLANETS MEDEA+) and FIT-330201-2004-11 (QUETZAL), by the Spanish Ministry of Science and Education TIC2002-04594-C02 and jointly financed by FEDER and by the Catalan Government SGR 2005-00996 and SGR2005-00690.

II. IMPULSE RADIO SIGNAL

In contrast to traditional narrowband systems where the carrier of information is a sinusoidal signal, in a communication link based on IR the carrier of information is a pulse train [2]. This section describes the main modulation and coding techniques for a generalized IR system. The exposition is organized in three parts. First, the signal model is introduced including different modulation schemes as well as the signal construction using spreading codes. Second, the expression of the PSD is provided. It is shown that the PSD of an IR modulated signal is composed of a continuous part that depends on the pulse spectrum and a discrete part consisting on high power spectral lines. These spectral lines are an undesirable effect that force the reduction of the transmission power to keep the radiated spectrum below the FCC mask. Finally, it is shown that the use of spreading codes, Time Hopping (TH) and Direct Sequence (DS) codes, reduce the power of the spectral lines, which results in a more efficient usage of the available power below the mask.

A. Impulse Radio Signal Construction

The IR signal is based on the transmission of a pulse train and can be described as follows,

$$s(t) = \sum_{i=\infty}^{-\infty} a_i p(t - T_i) \qquad (1)$$

where a_i is the pulse amplitude, T_i determines the pulse position and $p(t)$ is the pulse waveform. With *M-ary* modulation, *log2(M)* information bits are grouped to form one symbol. In *M-ary* Pulse Amplitude Modulation (M-PAM), a_i is chosen between the set {0, 1, …, M-1}, each representing one symbol value. In *M-ary* Pulse Position Modulation (M-PPM), the symbol pulse can be transmitted with M different delays, each representing one symbol value. In this case, the set of values for T_i is $\{p(t - m \cdot T_\Delta)\}$ where T_D is the so-called modulation index and $m \in \{0, \ 1, \ \cdots \ M-1\}$. Note that T_D must greater than the pulse width, which is denoted as T_P. Monocycle pulses are preferred because the radiation efficiency at the antenna is better [3]. For modern UWB-IR systems, the Gaussian monocycle is selected as a good candidate because it can be easily fitted below the FCC mask [3][4]. The mathematical expression in time and frequency domain of the Gaussian monocycle is given by,

$$p(t) = A \cdot \sqrt{2e}\,\frac{t}{t} \cdot e^{-\left(\frac{t}{t}\right)^2} \qquad (2)$$

$$P(f) = -jAp\sqrt{2pe} \cdot f \cdot t^{\,2} e^{-(pft)^2} \qquad (3)$$

where A is the pulse amplitude and t is the parameter that controls the pulse width, T_P. From (3), it follows that the central frequency f_c and the 3-dB bandwidth $BW_{3\text{-}dB}$ of the Gaussian monocycle are given by,

$$f_c = \frac{1}{tp\sqrt{2}} \qquad (4)$$

$$BW_{3-dB} = \frac{1.16}{tp\sqrt{2}} \qquad (5)$$

Let's define the central frequency of the pulse as the middle point between the limits of the FCC mask (3.1-10.6GHz). It follows that $f_c = 6.85GHz$ and thus, $t = 38.5\ ps$. Figure 1 depicts the time domain response of the Gaussian monocycle, resulting a pulse width of about $T_P \approx 180\ ps$. Figure 2 show the frequency domain of the same pulse. Note that filter requirements to meet the FCC regulations can be easily achieved.

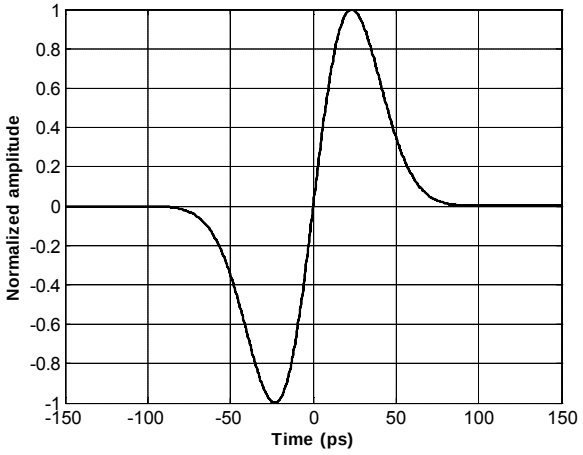

Fig. 1. Gaussian monocycle with $t = 38.5ps$.

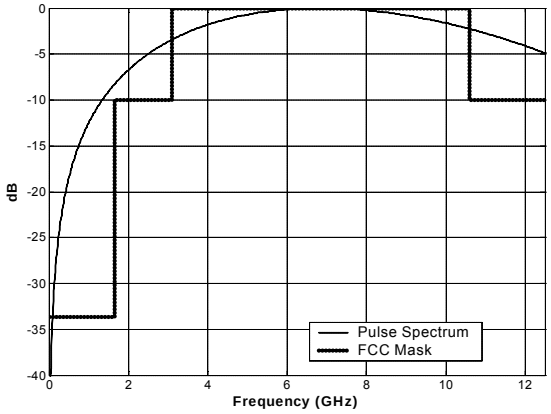

Fig. 2. Frequency response with $t = 38.5ps$.

IR signal are commonly based on spreading codes for several purposes. These codes can be considered to make the communication more robust against

interferences, or as a medium access mechanism [5]. The spreading codes can also be used to reduce the spectral lines in the PSD of the radiated signal, which is the main concern in an FCC compliant transmitter. When a spreading code is applied, more than one pulse is transmitted for each symbol.

Let's consider an IR communication system having one bit per symbol and working at a rate of $R_b = 1/T_b$ *bits/s*, where T_b is the bit period. In order to introduce the spreading codes, each bit interval is divided into *Ns* equally sized intervals of length *Ts*, named frame intervals. Note that $Ns \cdot Ts$ must be smaller than or equal to the bit period *Tb*. In the same way the frame interval is divided into N_h equally spaced time intervals known as chip intervals. In each frame interval only one modulated pulse is transmitted. The pulse sign or the position of the pulse inside the frame interval depends on the spreading code applied to the IR signal. Different spreading codes can be defined for generic IR signal:

·No Spreading Code. Only one modulated pulse is transmitted per bit. In this case $Ns = 1$, $N_h = 1$ and $Ts = Tb$.

·Repetition Code. The same modulated pulse is transmitted at the beginning of each frame interval. Note that both, pulse amplitude and pulse position are the same for all frame intervals of one bit. In this case $Ns > 1$, $N_h = 1$ and $Ns \cdot Ts \pounds Tb$.

·Direct Sequence Spread Spectrum Code (DS-SS). The modulated pulse is transmitted at the beginning of each frame interval. The sign of the pulse in each frame interval is modified according a direct sequence spreading code. In conventional DS-SS systems the code sequence is of length *Ns* and the same sequence is used for all bits. In this case $Ns > 1$, $N_h = 1$ and $Ns \cdot Ts \pounds Tb$.

·Time Hopping Spread Spectrum Code (TH-SS). In this case, for each frame interval only one pulse is transmitted on one of the chip intervals. The TH code determines the chip interval in which the modulated pulse will be located into each frame of each bit. N conventional TH-SS coding $Ns > 1$, $N_h > 1$ and $Ns \cdot Ts \pounds Tb$.

From a practical point of view the TH code is generated by a pseudo-random number generator, which inevitably results that the TH code will repeat itself. Let's denote T_{TH} as the repetition period and N_b the number of bits transmitted in this period. The signal transmitted in the k^{-th} period when TH and DS spread spectrum codes are used can be written as,

$$s_k(t) = \sum_{l=0}^{N_b} \sum_{h=0}^{N_s - 1} c_h^{DS,l} \cdot a_l^k \cdot p\left(t - lT_b - b_l^k T_\Delta - c_h^{TH,l} T_c - hT_s\right) \quad (6)$$

where $s_k(t)$ is the transmitted signal conveying the k^{-th} sequence N_b consecutive bits; a_{lk} is the amplitude of the *l-th* bit of the *k-th* sequence and it's used for M-PAM modulation. When no PAM modulation is used this parameter is set to $a_{lk} = 1$. b_{lk} is the time position of the *l-th* symbol of the *k-th* sequence. It carries the information corresponding to M-PPM modulation. When the PPM modulation is not used, this parameter is set to $b_{lk} = 0$. Inside each bit interval *h* denotes the frame number. $c_h^{DS,l}$ is the DS code chip amplitude for the *h-th* frame of the *l-th* symbol. When no DS code is used this parameter is set to $c_h^{DS,l} = 1$. $c_h^{TH,l}$ selects the chip interval of the *h-th* frame of the *l-th* symbol according with the TH code. When no TH code is used this parameter is set to $c_h^{TH,l} = 0$. T_s is the duration of the frame interval. Finally, T_c is the duration of the chip interval, T_b denotes the duration of the symbol interval and T_D represents the Modulation index for PPM.

As a matter of example consider an IR signal with the following parameters,

$$\begin{aligned} &N_b = 2, \quad N_s = 2, \quad N_h = 2, \quad 2 - PPM \\ &bit \;\; sequence \;\; \{b_0 = 1, b_1 = 0\} \\ &TH \;\; Code \;\; sequence \;\; \{0, \;\; 1, \;\; 1, \;\; 0\} \\ &DS \;\; Code \;\; sequence \;\; \{1, \;\; -1, \;\; 1, \;\; -1\} \end{aligned} \quad (7)$$

Fig. 3 shows the constructed signal corresponding to one period of the TH code. Note that the pulse sign is inverted according to the DS code whereas the pulse position depends on both, the TH code and the input bits.

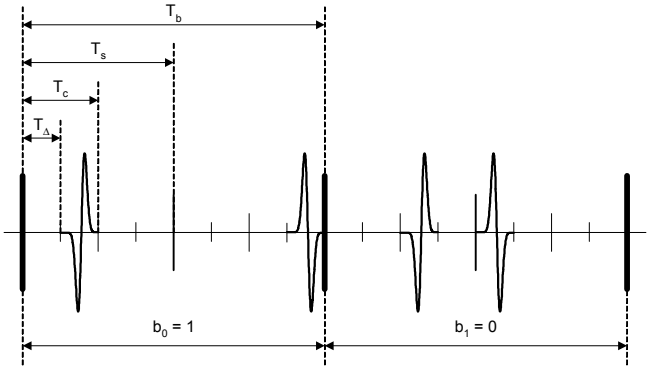

Fig. 3 Example of IR signal using TH and DS codes

B. Power Spectral Density

In general, the signal in (6) is a cyclo-stationary random process. As a consequence the PSD $S_y(f)$ consists of a continuous component $S_y^C(f)$ and a discrete component $S_y^D(f)$ [6]. The continuous part is mainly dependent on the pulse spectrum. The discrete part is a set of spectral lines, whose separation in frequency is

proportional to the periodicity of the cyclo-stationary process. In some IR systems these spectral lines could be exploited for timing synchronization at the receiver [7]. However, in an FCC compliant transmitter they can be problematic because they cause interference to narrow-band systems. A possible solution to the latter problem could be to reduce the transmitted power to assure that all spectral lines are below the FCC mask, but in this case the total transmitted power would be too small for a reliable link. Therefore, we can define the spectrum efficiency as follows,

$$h = \frac{Peak\ Power\ of\ Discrete\ Part}{Maximum\ Power\ of\ Continuous\ Part} \qquad (8)$$

Analytical expressions of the PSD for a generalized UWB-IR signal in (6) can be found in [8] and [1]. In [8] the PSD is calculated taking into account the effect of the timing jitter and assuming a completely random TH code. A PSD expression is given in [1] for both, short and long TH spreading codes. Following the notation in [1], the PSD of a generalized signal described in (6) can be written as:

$$P_y(w) = \frac{1}{T_{TH}} E\left\{\left|S_p(w)\right|^2 - E\left\{S_p(w) S_q^*(w)\right\}\right\} + \frac{1}{T_{TH}^2} E\left\{S_p(w) S_q^*(w)\right\} \sum_k d\left(w - \frac{2p\,k}{T_{TH}}\right) \qquad (9)$$

where p and q are used to identify two different waveforms signals conveying N_b symbols, T_{TH} is the duration of signal $s_p(t)$ and $S_p(w)$ if the Fourier Transform (FT) of $s_p(t)$. The continuous and discrete part of the PSD can be clearly distinguished as the first and second term respectively of (9).

Let's consider the case when 2-PPM using TH-SS and DS-SS codes is used. In this case the period of the TH code is a number N_b of consecutive symbols. Formulating as in [1] the PSD can be written as,

$$S_y(f) = \frac{\left|P(f)\right|^2}{T_{TH}} \left\{\left\{R_0^a - R_1^a R_1^b(w)\right\} \sum_l T_l(w) T_l^*(w)\right\} + \frac{\left|P(f)\right|^2}{T_{TH}^2} \left\{R_1^a R_1^b(w) \sum_{l,n} T_l(w) T_n^*(w)\right\} \Pi_{TH}(w) \qquad (10)$$

where $\left|P(f)\right|^2$ is the spectrum of the pulse,

$R_0^a = E\left\{a_l^p a_n^p\right\}$ for $l=n$,

$R_1^a = E\left\{a_l^p a_n^p\right\}$ for $l \neq n$,

$R_0^b(w) = 1$ for $l=n$,

$R_1^b(w) = E\left\{e^{-jw\left(b_l^p - b_n^p\right)T_\Delta}\right\}$ for $l \neq n$,

$T_l(w) = \sum_h e^{-jwc_{l,h}T_c} e^{-jwhT_f} e^{-jwlT_b}$,

$\Pi_T(w) = \frac{1}{T} \sum_k d\left(w - \frac{2p\,k}{T}\right)$.

Note that with the term $T_l(w)$, which depends on the TH code, and the term R_1^a, which depends on the DS code, the discrete part of the PSD can be reduced, improving the spectrum efficiency.

C. Time Hopping Codes Generation

There are a number of code generators that might be used to generate the TH codes for UWB systems [2] [9]. The main purpose for using TH and DS codes in this IR transmitter is to reduce the spectral lines of the radiated PSD. However, it can be shown that there will be spikes at the spectrum that cannot be removed by using spread spectrum codes [2]. Maximum Length (ML) codes present a good compromise between complexity of implementation and the spectrum efficiency.

The TH code is a sequence of integers inside the set $\{0, \ldots N_S\text{-}1\}$. The ML codes for time hopping can be generated as shown in Fig. 4. It is a pseudo-random bit generator based on a primitive polynomial. We have chosen the following primitive polynomial to generate the TH codes,

$$P(x) = 1 + x^2 + x^{10} \qquad (11)$$

The Symbols of the TH codes can be obtained by multiplying some bits of the register counter by powers of two as expressed as follows,

$$S = \begin{bmatrix} s_0 & s_1 & \cdots & s_M \end{bmatrix}^T$$
$$TH_S = \begin{bmatrix} 2^0 & 2^1 & 2^2 & \ldots & 2^n \end{bmatrix} \bullet b \qquad (12)$$

where $M = log_2(N_S)$ and TH_S is a vector containing the sequence of the TH code.

As a matter of example let's consider an IR transmitter with the following parameters:

- Pulse waveform: Gaussian Monocycle
- Number of frames per symbol $N_S = 8$.
- Number of chip Intervals per frame, $N_h = 16$ ($M = 4$). A 4-bit TH code is generated
- Number of symbols in a period of the TH code, $N_b = 32$
- $T_c = 25ns$
- $T_D = 1ns$
- $T_{TH} = 102.4\ ms$
- Taking one bit of the scrambler of Figure 5 is formed the DS code.

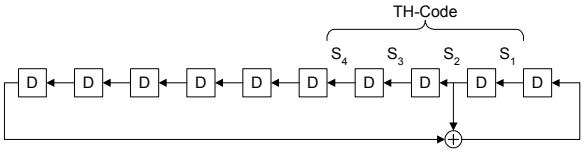

Fig. 4. Generation of ML Spreading Codes.

Fig. 5 shows the resulting PSD. The ratio between discrete part of the spectrum and the continuous part is reduced to approximately 10 dB.

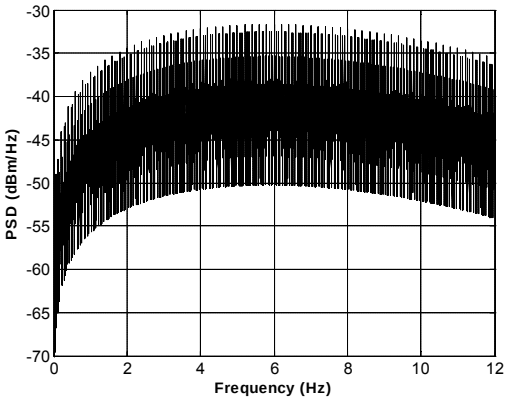

Fig. 5. PSD using DS and TH Spreading codes.

III. TRANSMITTER ARCHITECTURE

The IR transmitter is composed of a digital unit, in which the baseband processing is executed, a Digital Pulse Amplifier (DPA), a dual channel pulse generator, an Inverter/Non-Inverter block, which inverts the sign of only one of the pulse channels and a power combiner to radiate both channels on the same antenna. Fig. 6 shows the block diagram of the transmitter.

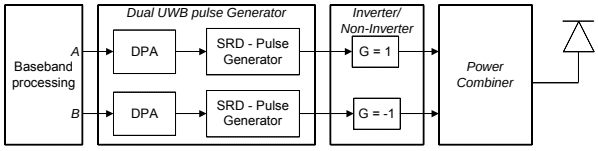

Fig. 6. Transmitter Block Diagram

The digital unit constructs the IR signal using TH codes as described in the previous section and delivers the sequence of digital pulses. The DS code is implemented by generating those pulses with DS code equal to 1 through channel A and those pulses with DS code equal to –1 through channel B. The two DPA drive the digital pulses to the dual channel pulse generator. The pulse generator is based on Step Recovery Diodes (SRD). As a result, the transmitter is able to generate two Gaussian monocycle pulses with opposite polarity for DS-SS codes.

A. Digital Pulse Amplifier

The digital pulse amplifier consists of a Class-S amplifier, which drives a medium power bipolar transistor to work as a switch [10]. The bipolar transistor is the responsible of delivering enough current to the SRD pulse generator. Pulse amplifier circuit is depicted in Fig. 7. Transistors Q1 and Q2 (Infineon BFT92P and BFR92 respectively) perform the two-position switch. The complementary pair Q1 and Q2 ensures rapid switching and adequate voltage to maintain the ON/OFF state of the output transistor Q3 (Infineon BFG235). The zener diode (Z_1) and base resistance (R1 and R2) are used for accommodating the voltage levels in DC mode. In order to raise the switching speed base capacitors C1 and C2 are introduced. The voltage variation through those capacitors generates additional current, which contributes to the rapid removal of stored charge in the PN-junction of the transistors. The low pass filter formed by the RC circuit limits the maximum Pulse Repetition Frequency (PRF). The constant time of this RC circuit must be $RC < 1/10 f_p$, where f_p is the maximum pulse repetition frequency. It must be noticed that the SRD pulse generator requires 50 Ohms output impedance of the pulse amplifier.

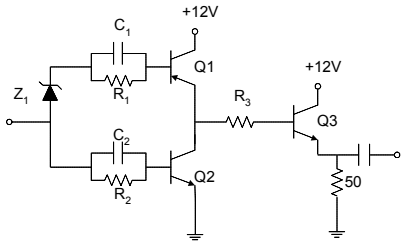

Fig. 7. Digital Pulse Amplifier

B. SRD Pulse Generator

SRD diodes have been used during years for the generation of short duration pulses [11]. The PRF of conventional SRD is often about 10 MHz. On the other hand, in UWB systems the PRF is needed to be around 60 MHz, combined with 3 V pulses, to obtain the maximum PSD allowed level. In this section we present a transmitter capable to operate up to 40 MHz thanks to a high-speed digital pulse amplifier able to drive enough current for the SRD circuit.

The main characteristic of a SRD diode is the very fast switching from forward to reverse modes. This property is the one that permits to generate falling or raising edge signal of tenths of picoseconds. While the diode is forwarded- polarized electric charge is stored in the P-N junction of the diode. The total stored charge depends on the average current and the recombination time of the diode according to

$$Q_F = I_F \cdot t \qquad (13)$$

where Q_F is the stored charge in the forward mode I_F is the average current flowing through the PN-junction and t is the recombination time of the diode.

Fig. 8 shows the designed Gaussian pulse generator circuit. It is based on typical circuit configurations for Gaussian pulse sharpening using SRD diodes [11] and monocycle Gaussian circuits [12] [13]. At the beginning of operation, the voltage source V_{in} is set to ground level. In this mode the SRD-1 operates in forward polarization through the bias circuit. The total stored charge in the PN-junction is given by (13). When the voltage source V_{in} rises (it can take a few nanoseconds), reverse current flows through the diode removing the stored charge. During this discharge period the diode keeps the low impedance state and consequently its voltage remains close to ground level. When the stored charge is completely removed, the SRD-1 switches abruptly to a high-impedance state. In this moment the input voltage is transmitted to the load. It generates a rising edge which duration is equal to the switching time of the diode, that can be lower than 100ps.

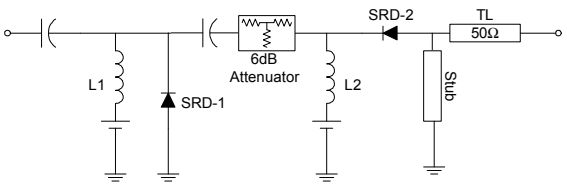

Fig. 8. SRD Gaussian pulse generation circuit

In the same way that SRD-1 is used to configure the rising edge of the pulse, SRD-2 is used to generate the falling edge of the Gaussian pulse. Whereas SRD-1 is operating on low impedance state, the output of the first state does not change from low level. During this time SRD-2 also operates in forward polarization through the bias circuit. Again the stored charge of SRD-2 is given by (13). In the moment in which SRD-1 switches to high impedance state, the rising edge is propagated to SRD-2. During the rise time, reverse current flows through SRD-2 removing the stored charge. When the PN-Junction of SRD-2 is completely empty of charge, it switches to high impedance state forcing the falling edge of the Gaussian pulse.

Following the Gaussian pulse generator, a passive shaping- network is used to generate a Gaussian monocycle pulse. Different shaping-network topologies may be used to perform this operation [11][12]. Short-circuited stub is a simple an easy network that works properly to create the Gaussian monocycle pulse. The stub generates a reflection of the pulse inverting the phase and introducing a delay t_P that depends on its length. This reflection, added to the former pulse, produces the desired pulse. To obtain the optimum Gaussian monocycle, t_P should be the half of the original pulse generated.

In order to increase the amplitude of the pulse and reduce the ripple and/or reflections generated by the circuit, some modifications have been done to the typical circuit. The 50-Ohm load used in [11] after SRD-2 has been eliminated and SRD-2 is then connected directly to ground using the short-circuited stub of the shaping network. Then, a reflection still appears at the output because the pulse goes through de SRD-2 backwards to SRD-1 generating again another pulse. To solve this problem a 6dB-attenuator has been placed between the diodes, so that the reflection is then considered negligible. Although this solution reduces the amplitude, the pulse generated has improved ripple as depicted in Fig. 9.

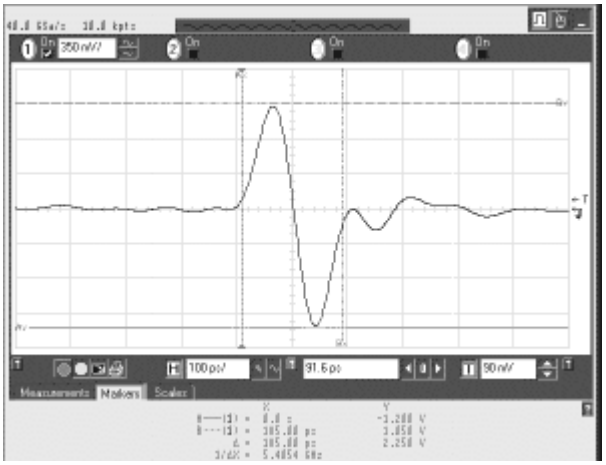

Fig. 9. Gaussian Monocycle Pulse (185 ps @ 2.2 V)

The Gaussian Monocycle pulse is not completely symmetric (87,5%). Inductance at SRD pins is the main fact that contributes to the asymmetry of the pulse [11]. Other characteristic that also affects is that the fall and rise of the pulse (before the shaping network) are produced in different ways. Whereas the rise is produced by a slow edge done by the digital amplifier towards SRD-1, the fall is consequence of the fast transition of the SRD-1 diode going through SRD-2.

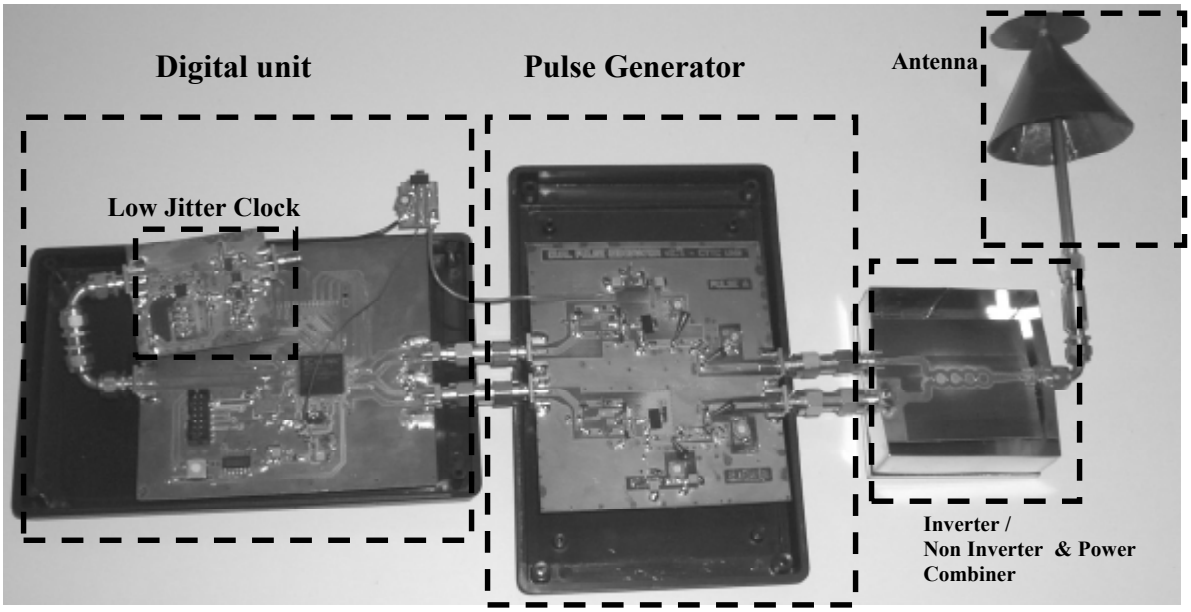

Fig. 10 Impulse Radio Transmitter

C. Inverter / Non- Inverter Circuit

The inverter key point lies in the transitions from unbalanced to balanced line. Changing the output ground plane side causes the variation of the electric potential reference, and the consequent pulse inversion.

The use of active devices to perform the inverter is refused due to their limited bandwidth. An alternative passive inverter circuit using slotlines is studied in [14]. This technique is also discarded due to the narrower slots needed to achieve low impedance slotlines. Passive pulse inverting technique based on Double-Sided Parallel-Strip Lines combined with microstrip lines has been chosen.

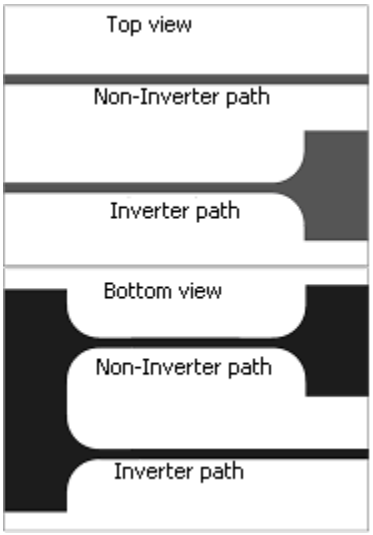

Fig. 11. Inverter / Non Inerter Scheme

A 50Ohms Double-sided parallel-strip line connects the input and output ports of the inverter. Transitions between the Double-sided parallel-strip lines (balanced lines) and the microstrip lines (unbalanced) are needed. Different types of these transitions (called baluns) are analyzed in [15]. Circularly tapered transitions are used in the presented design.

The Inverter/Non-Inverter circuit has two different paths: one performs the pulse inversion whereas the other does not change the pulse sign. In Fig. 11 is depicted the final layout of this stage including the top and bottom layers. The Non-Inverter branch line has the same length as the Inverter branch. This is necessary to maintain both ways synchronized and same strip loses.

D. Power Combiner

A Wilkinson Power Divider is used to combine the inverted and the non-inverted pulses. Wilkinson structure has been chosen due to good matching in all ports and good isolation between the combined ports. To increase the useful bandwidth a Four-Stage Wilkinson Power Combiner is designed [16].

A complete circuit composed of the "Inverter/Non-Inverter" and the "Wilkinson Combiner" has been built. In order to provide a direct connection to the foreseen Double-Sided strip antenna, the Wilkinson Combiner is also based on Double-sided parallel-strip lines.

The characteristic impedance of a double-sided parallel-strip line with dielectric separation h is twice the characteristic impedance of a microstrip line with dielectric thickness $h/2$ (with the same strip width) [15]. Resistance values used in double-sided parallel-strip

Wilkinson combiner also changed with reference to former microstrip configuration. The used structure uses 50 Ohms instead of 100 Ohms placed in microstrip.

E. UWB Antenna

In order to make measurements we have implemented a discone antenna due to its omni-directional radiation property and broadband characteristic. Fig. 12 shows the main the mechanical diagram of the developed antenna.

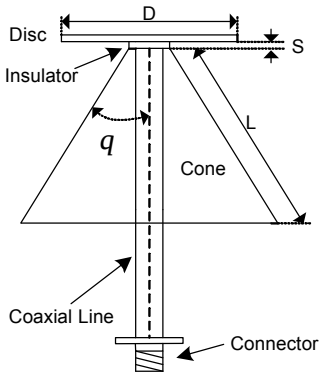

Fig. 12 Discone Antenna Diagram

To calculate the dimensions we used [17][18],

$$L(inch) = \frac{2958}{F}(13) \quad q = 25^{\circ} - 40^{\circ} \tag{14}$$

$$D(inch) = \frac{2008}{F} \tag{15}$$

where F is the starting frequency in *MHz*. The cone length L and disc diameter D fix the starting frequency of the antenna. To achieve the maximum frequency range the separation between the disc and the cone S should be as short as possible.

In Fig. 13 is shown the input return loss measured at the network analyzer. The parameter S11 is kept below *10 dB* from *1.5GHz* to frequencies above *11GHz*.

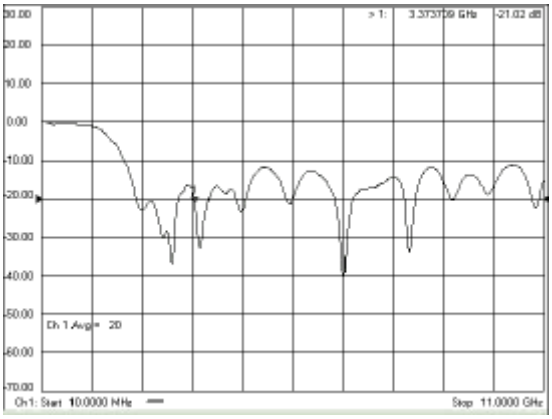

Fig. 13. Antenna Input Return Loss (dB)

F. Reference Oscillator

Considering that UWB pulses are about 200 ps width, the reference oscillator clock that controls the transmission or the reception must have very low jitter or, in other terms, very low phase-noise. If this fact is not achieved synchronization between the transmitter and the receiver will be highly degraded.

A low jitter clock has been implemented using Maxim MAX2620 low phase noise oscillator (-110dBc/Hz) and Hittite HMC394LP4 frequency divider. The oscillator can work at frequencies up to *1 GHz*. The divider consists on a 5 bit programmable digital counter, which provides lower frequency clock signals. Fig. 14 shows jitter measurement of the developed reference oscillator. The result gives a rms-jitter of *1.7 ps* and a peak-to-peak jitter of *10 ps*.

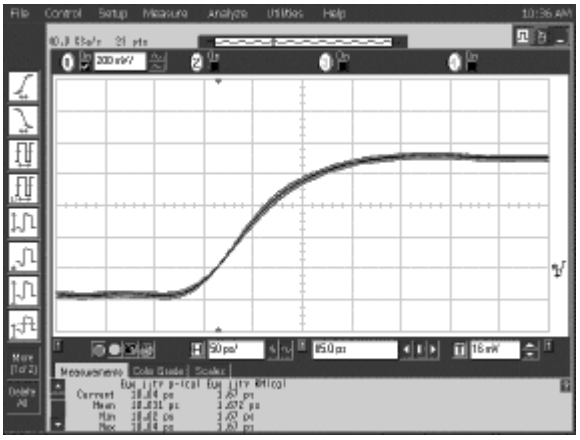

Fig. 14. Jitter measurement

IV. MEASUREMENT RESULTS

The whole transmitter, as shown in Fig. 10, has been fabricated in microstrip technology using a substrate TACONIC RF30-0300, having *0.76mm* thickness and a relative permittivity e_r *=3*. The SRD diodes by Micrometrics (MSD700-19-1) have a switching time of about *70ps*.

Fig. 15. Combined pulses at the same port..

The implemented transmitter generates two Gaussian monocycle pulses. The CPLD generates pulses through two output ports that are connected to their respective SRD pulse generator. Once the two pulses are generated, the Inverter/ Non-Inverter block produce a positive and a negative pulse needed when using DS-SS codes. Finally, the combiner permits to transmit the pulses using the same antenna. Fig. 15 measurement result when both, the inverted and the non-inverted pulses are generated through the common output port.

Using TH-SS codes allows to reduce the peak-to-average ratio in the PSD, or the spectrum efficiency as defined in (8). Fig. 16 shows the measurement result when TH and DS spreading codes are used. The spectrum efficiency has been reduced to *10 dB* as the theoretical result of Fig. 5 shows. The maximum output level is about *–45dBm/Mhz*, which is close to the permitted limit by FCC mask (*–42dBm/Mhz*). Finally note that total 10dB bandwidth of the pulse goes from 2 to 9 GHz. This corresponds to a Gaussian monocycle pulse having $t = 50ps$ and a pulse width of about *185ps*.

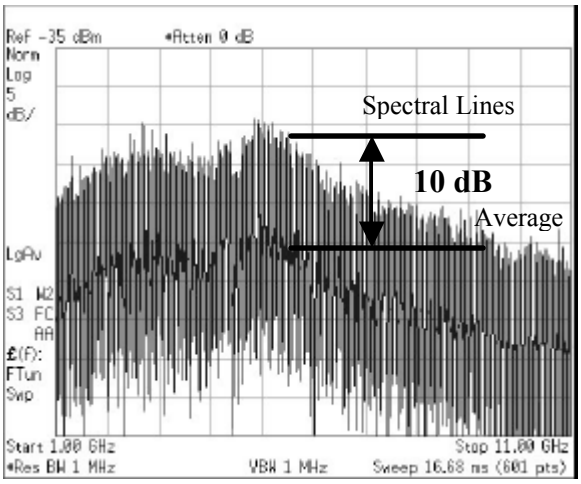

Fig. 16. Generated PSD using TH – DS Codes

V. CONCLUSIONS

This paper presents an UWB transmitter based on impulse radio, which is a hot research topic since the FCC delivered the regulations regarding UWB in 2002. The main concern in systems compliant with the FCC mask is the efficient usage of the available power below the mask limitations. To this end, the use time hopping and direct sequence spreading codes are used to reduce the level of the spectral lines in the radiated PSD. Maximum length sequences for TH codes are presented as a good compromise between complexity and the achieved spectral efficiency.

The developed transmitter uses an SRD-based circuit and a passive shaping network to generate a Gaussian monocycle. The measured pulse width is about 185ps. A Class-S digital pulse amplifier using conventional RF transistors has been designed to drive SRD pulse generator that can work up to 40MHz. To implement the DS spreading codes, a pulse inversion block has been developed using UWB transitions between parallel doubled sided strip lines and microstrip lines, shown an ideal performance for this application. The MAXIM 2620 shows very good Jitter performance for UWB systems, which presents an rms-jitter of *1.7ps*. With respect to the antenna, a discone antenna is developed to perform field measurements. This antenna is operative from 1.5GHz up to 11GHz.

Measurements on the generated PSD show a maximum level of –45dBm/MHz, which is kept below the mask limitation at –42dBm/MHz. In addition the measured PSD using TH and DS spreading codes corresponds to theory predictions giving a ratio between the peak power of the spectral lines and the pulse spectrum of about 10dB.

REFERENCES

[1] J. Romme, L. Piazzo, "On the Power Spectral Density of Time-Hopping Impulse Radio", Ultra-Wideband Systems and Technologies, 2002, Digest of papers 2002, IEEE Conf. on, May. 2002.

[2] C. Müller, S. Zeisberg, H. Seidel, A. Finger, "Spreading Properties of Time Hopping Codes in Ultra Wideband Systems", IEEE 7º Symp. On Spread-Spectrum Tech. & Appl., Prague, Czech Republic, Sept. 2002.

[3] X. Chen, S. Kiaei, "Monocycle shapes for Ultra Wideband System" Proceedings IEEE 2002.

[4] Sheng, P. Orlik, A.M. Haimovich, L. J. Cimini, Jr. J. Zhang, "On the spectral and Power Requirements for Ultra-wideband Transmission", Proceedings IEEE 2003

[5] M.Z. Win, R.A. Scholtz, "Impulse Radio: How It Works", IEEE Commun. Letters, Vol. 2, No. 2, Feb. 1998

[6] S. Benedetto, E. Biglieri, *Principles of Digital Transmission With Wireless Applications*, Kluwer Academic / Plenum Publishers, New York 1999.

[7] O. Albert, C.F. Mecklenbräuker, "Low-Power Ultra-Wideband Radio Testbed for Short Range Data Transmission", Int. Workshop on Ultra-Wideband systems (IWUWBS 2003), June 2003.

[8] M.Z. Win, "A unified Analysis of Generalized Time-Hopping Spread-Spectrum Signals in the Presence of Timing Jitter", IEEE Journal on Selected Areas in Communications, Vol. 20, No. 9, Dec. 2002.

[9] S.V. Maric, E.L. Titlebaum, "A Class of Frequency Hop Codes with Nearly Ideal Characteristics for Use in Multiple-Access Spread Spectrum

Communications and Radar and Sonar systems", IEEE Trans. On Commun., Vol. 40, No. 9, Sept. 1992.
[10] H. L. Krauss, Ch. W. Bostian, F. H. Raab, "Solid State Engineering", John Wiley & sons, Canada 1980.
[11] "Pulse and Waveform Generation with Step Recovery Diodes", Hewlett Packard AN-918.
[12] J.S. Lee, C. Nguyen, T. Scullion, "New Uniplanar Monocycle Pulse Generator and Transformer for Time-domain Microwave Applications", IEEE Trans. On Microwave Theory and Techniques, Vol. 49, No. 6, June 2001.
[13] A. Mollfulleda, P. Miskovsky, J. Mateu, "Rubost Shaping Network for Impulse Radio and UWB Signal Generator", 2005 IEEE MTT-S International Microwave symposium, 12-17 June 2005.
[14] K. C. Gupta, R. Garg, I.J. Bahl, and P. Bhartia, *Microstrip Lines and Slotlines*, 2nd ed.Artech House, Inc., 1996.
[15] S.G, Kim, K. Chang, "Ultrawide-Band Transitions and New Microwave Components Using Double-Sided Parallel-Strip Lines", IEEE Trans. On Microwave Theory and Tech., vol. 52, No. 9, September 2004.
[16] D.M. Pozar, "Microwave Engineerin, 2nd Edition", 1998, John Wiley & Sons, Inc.
[17] Matthew R. Wills , A Different Antenna for the Mode-Stirred Chamber , http://www.ieee.org/organizations/pubs/newsletters/emcs/fall01/wills.html
[18] A UHF Discone Antenna for scanners, http://www.northcountryradio.com/Articles/discone.htm

Signal Processing and Spectrum Use

Brett Glass
Advanced Communication Concepts
Phone: (307) 761-2895
Email: brett@lariat.org

Paper unavailable at time of printing.

Tri-Band RF Transceivers for Dynamic Spectrum Access

Nishant Kumar, Yu-Dong Yao, Terrance Bernard, and Amitoj Singh

Wireless Information Systems Engineering Laboratory (WISELAB)
Department of Electrical and Computer Engineering
Stevens Institute of Technology, Hoboken, NJ 07030
Email: {nkumar, yyao}@stevens.edu

Abstract: *In order support dynamic spectrum access research and conduct DSA related experimentations, we have developed an RF transceiver test bed which operates over three industrial, scientific and medical (ISM) bands (902-928 MHz, 2.4-2.4835 GHz, and 5.725-5.850 GHz). The paper describes an approach where a baseband signal is up-converted at the transmitter to the ISM band and then down-converted to the base band at the receiver. The paper shows the basic architecture of the design and concludes with some results of an experiment done using this test bed.*

1. Introduction

In a recent spectrum reform policy document [1], four basic spectrum usage models were considered, (a) government planning and close supervision of spectrum, (b) spectrum as private property, (c) spectrum access markets, and (d) spectrum commons. In order to improve spectrum utilization, the concept of dynamic spectrum access (DSA) [1, 2] was proposed by FCC Spectrum Policy Task Force. Key functions in implementing DSA include sensing the radio frequency (RF) environment, identifying available channel resources, and adapting transmissions through the available channels. We focus our research on the DSA RF environments, specifically, the interference and noise characteristics, and developing access and resource allocation algorithms for DSA systems. In order to support our ongoing research and conduct DSA related experimentations, we have developed an RF transceiver test bed which operates over three industrial, scientific and medical (ISM) bands (902-928 MHz, 2.4-2.4835 GHz, and 5.725-5.850 GHz).

The outcrop of the recent wireless devices can be attributed to ISM bands which were initially reserved internationally for the non-commercial and unlicensed use of RF electromagnetic fields for industrial, scientific and medical purposes. They are unlicensed but are governed by rules made with respect to the transmission technology and power. These rules have helped in the congenial growth of short range devices and made helped in the development of technologies which are not only power efficient but has also increased the density of transmitting unit to an unprecedented number. The FCC regulation not only allowed non-governmental uses of these frequencies, but also the unquestioned use of radio spectrum giving us the freedom to experiment new technologies and ideas.

There are three commonly used ISM bands, 902 – 928 MHz, 2.4 – 2.4835 GHz and 5.725 – 5.850 GHz. FCC specifies that all the transmission in these bands should be limited to frequency hopping and direct sequence spread spectrum and the peak power output of the transmitter shall not exceed 1 Watt. If transmitting antennas of directional gain greater than 6 dBi are used, the power shall be reduced by the amount in dB that the directional gain of the antenna exceeds 6 dBi. Frequency hopping systems shall have hopping channel carrier frequencies separated by a minimum of 25 kHz or the 20 dB bandwidth of the hopping channel, whichever is greater.

This paper introduces a test-bed targeted to exploit the three ISM bands. With the use of off-the-shelf components and simple design, transmission and reception can be done at any desired frequency in the ISM band. Easy control and simple interface allow it to be integrated with computers and support in the development of communication protocols, especially, for dynamic spectrum access systems. This setup can also be used for measurement of channel usage and noise level. A fast sweep through a target band of frequency and subsequently storing the data for further analysis can be easily archived.

2. Architecture and Descriptions

Figure 1 shows the block diagram for the first band (902 – 928 MHz) transmission and reception.

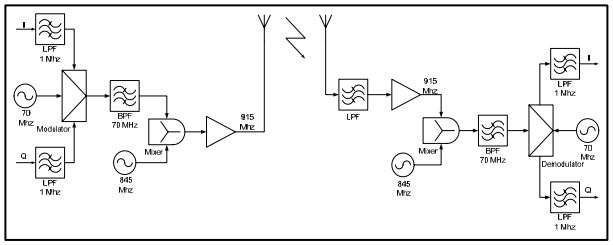

Fig. 1. Block diagram for 902 – 928 MHz band operation.

In the following, we describe the architecture and components in implementing the 902 – 928 MHz transceiver.

Transmitter: The basic components of the transmitter of the transceiver are the modulator, band-pass filter, mixer, local oscillator (LO) source and amplifier. The design also uses low pass filters at the IQ end to remove noise from the input base-band signal. The signal is quadrature modulated at 70 MHz. The 70 MHz needed for the modulation is provided by a frequency synthesizer. The modulated signal is passed through a band pass filter with centre frequency of 70 MHz. This removes noise from the modulated signal. The filtered signal is then fed into a mixer, which accepts the signal at the IF end and a frequency of 845 MHz by the frequency source at the local oscillator input and produces a carrier signal at 915 MHz. This signal is amplified and transmitted.

Receiver: The receiver end of the board reverses the operation of the transmitter. The received signal is first sent through a low-pass filter to remove most of the noise/interference. This signal is then fed into an amplifier before being sent into a mixer which receives a frequency of 845 MHz from the local oscillator and subtracts this from the 913 MHz signal input to have a signal of 70 MHz generated at the output. A band pass filter is used to filter this signal out and send the signal to the demodulator. The demodulator takes a 70 MHz input from the local oscillator to demodulate the signal which recovers the original signal.

Similar design is used for the other two ISM bands (2.4 – 2.4835 GHz and 5.725 – 5.850 GHz) with switches to switch over different frequency bands. Signal can be tapped at any point in the circuit and can be analyzed using an oscilloscope or a spectrum analyzer. Figure 2 shows a high-level block diagram for the tri-band operation.

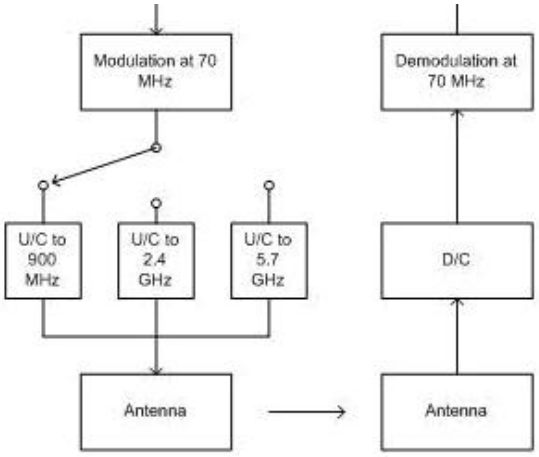

Fig. 2. A high-level block diagram for tri-band operation.

Figure 3 shows a developed prototype of the tri-band transceivers.

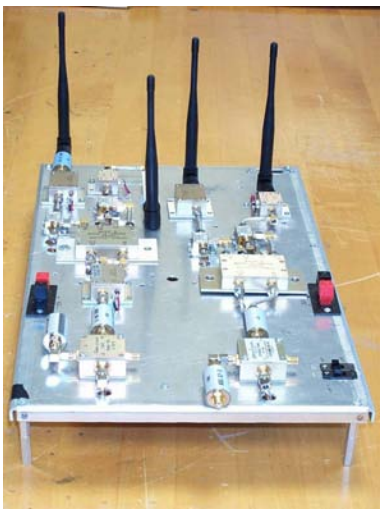

Fig. 3. A tri-band prototype.

3. Experiment

We have been using the developed tri-band for dynamic spectrum access investigation. The test bed is used to measure the channel and interference. The test bed facilitates in targeting a band of channel and sweeping through to measure the strength of signal. This can be done at baseband frequency or at the intermediate frequency.

Figure 4 (a) through (f) shows photos taken during a transmission/reception experiment.

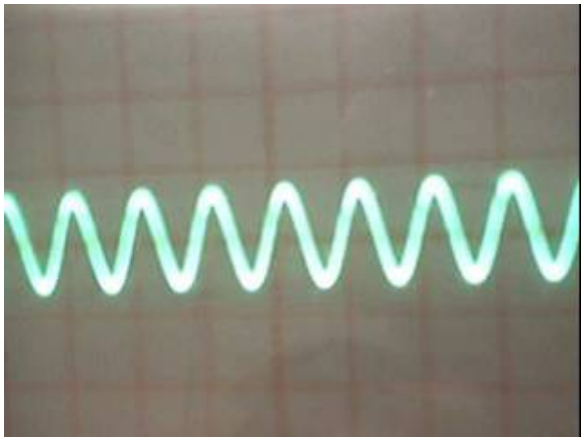

Fig. 4 (a). Input signal.

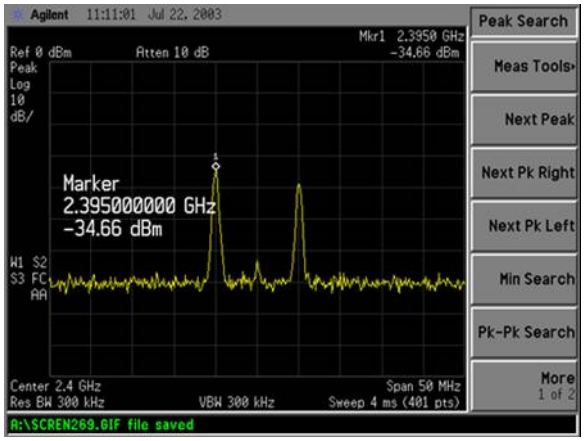

Fig. 4 (d). Signal transmitted.

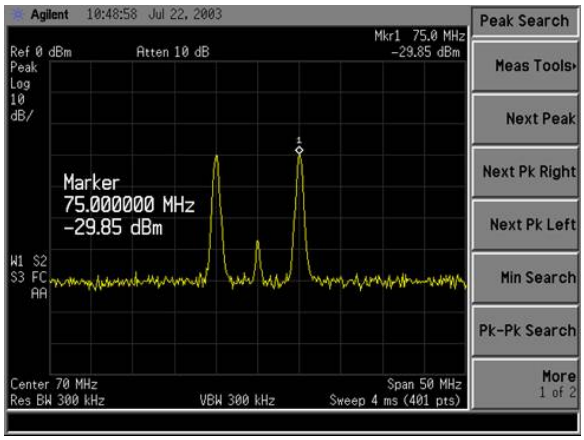

Fig. 4 (b). Signal modulated to 70 MHz.

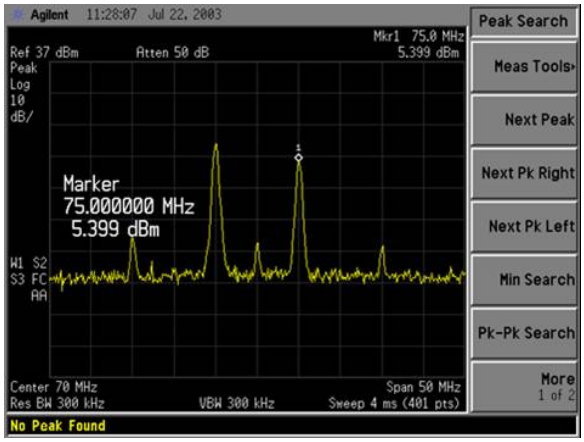

Fig. 4 (e). Signal down-converted to 70 MHz.

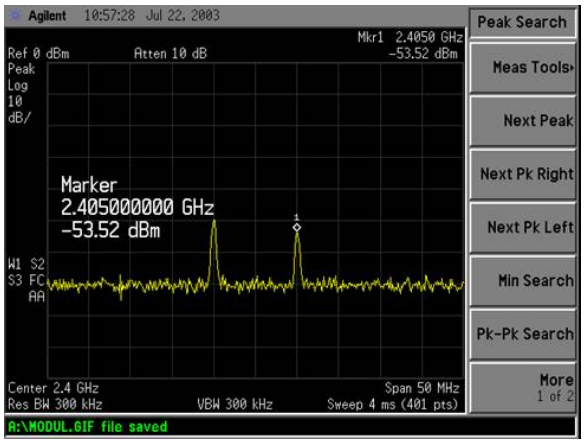

Fig. 4 (c). Signal up-converted to 2.4 GHz.

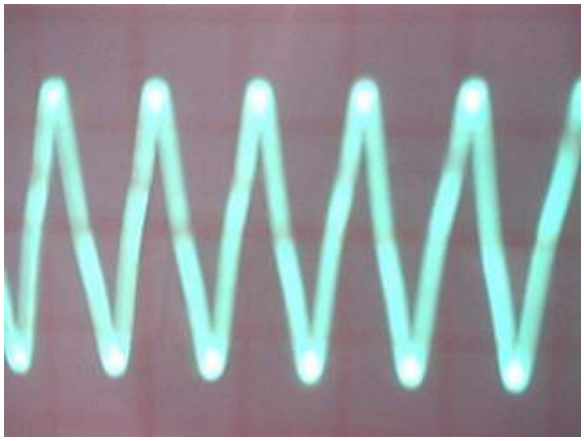

Fig. 4 (f). Output signal.

The above photos show the basic transceiver flow in a communication system where it starts with a baseband input signal, Fig. 4 (a), which is modulated by an intermediate frequency, Fig. 4 (b). This signal is then up-converted to 2.4 GHz frequency, Fig. 4 (c), which is amplified and transmitted, Fig. 4 (d). At the receiver side, the signal is filtered and then down-converted to the intermediate frequency, Fig. 4 (e), and finally, demodulated to baseband, Fig. 4 (f).

4. Conclusions

This paper first discusses the use of a tri-band RF test bed for dynamic spectrum access research. It describes the architecture of the test bed and its elements. The transmitter and receiver block diagrams and circuits are shown and the designs are discussed. An experimentation using the test bed using the test bed is also presented.

Some features of the developed test bed include flexibility to perform testing at every frequency stage and the components/designs can be changed according to the requirements of testing scenarios. The minimum detectable signal (MDS) is -62 dBm as compared to -52 dBm for typical off-the-shelf transceiver boards. The local oscillator stability is in the range between +/- 1 ppm and +/- 2.5 ppm as compared to +/- 75 ppm for typical off-the-shelf transceiver boards.

References:

[1] R. M. Entman, *Balancing Policy Options in a Turbulent Telecommunications Market: A Report of the Seventeenth Annual Aspen Institute Conference on Telecommunications Policy*, The Aspen Institute, 2002.

[2] P. Kolodzy, *Next Generation Spectrum Policy*, Spring Meeting of the Aspen Institute Conference on Telecommunications Policy, April 2002.

[3] *Frequency Spectrum Rules and Regulations,* FCC Rules: CFR Title 47 Part 18. (http://www.airlinx.com).

[4] B. Razavi, *RF Microelectronics*, Prentice Hall PTR, November 1997.

[5] H. Taub and D. L. Schilling, *Principles of Communications Systems*, McGraw-Hill Companies, 2nd edition, April 1986.

[6] *Mini-Circuits Component Datasheets*, (http://www.minicircuits.com).

[7] D. Grandblaise, D. Bourse, K. Moessner, and P. Leaves, "Dynamic spectrum allocation in composite reconfigurable wireless networks," *IEEE Communications Magazine*, vol. 42, no. 5, pp. 72-81, May 2004.

[8] V. Brik, E. Rozner, S. Banerjee, and P. Bahl, "DSAP: A Protocol for Coordinated Spectrum Access," *First IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN)*, pp. 611 – 614, November 2005.

Multimedia Quality of Service and Net Neutrality on Wireless Networks

Tom Lookabaugh and Douglas C. Sicker

Interdisciplinary Telecommunications Program and Department of Computer Science
University of Colorado at Boulder
[Tom.Lookabaugh, Douglas.Sicker]@colorado.edu
(303) 492-0429
Fax (303) 492-1112

Multimedia traffic differs in its demand on networks from traditional internet access applications, in particular requiring sustained sessions and high rates. A shift towards multimedia traffic consequently has a direct impact on the amount of oversubscription – or sharing of capacity among users - that is acceptable on a given network. Oversubscription is disproportionally important to the economics of wireless internet access providers relative to other access technologies such as cable and DSL. This means that wireless providers can be expected to feel more strongly the economic rationales for discrimination among content based on traffic type. But such discrimination flies in the face of the principle of net neutrality – a social policy goal that attempts to minimize discrimination in order to encourage innovation and openness on the internet. The consequence is that we can expect wireless internet access to be at the center of the debate about reasonable policy approaches to network neutrality that attempt to separate more and less socially desirable economic rationales for discrimination.

1. Introduction

Net neutrality – the principle that internet access providers should not discriminate among different instances of internet traffic – is at the core of an important policy debate about how the internet will evolve. Proponents of net neutrality claim it is indispensable to the innovation and openness of the internet that society has come to expect and cherish. Skeptics argue that service providers either have every right to discriminate among traffic types since they own the networks (an argument that then runs into serious questions of what society wishes to permit in tradeoffs between property rights and social policy goals) or that there are simple and compelling economic reasons for discrimination among traffic types based on their consumption of network resources.

The latter argument – that of compelling economic rationales for discrimination – is sharp in the case of a shift from traditional internet access applications such as web browsing towards an increasing fraction of multimedia traffic such as voice, music, and video. Moreover, we argue here that wireless internet access providers are disproportionally affected by this shift relative to other service providers such as DSL or cable modem. The consequence is that wireless internet access providers are likely to find themselves among the first attempting to defend deviations from net neutrality for economic reasons and so constitute a key test case and audience for any policy proposals that seek to balance, encourage, or enforce the reasonable application of net neutrality.

2. Multimedia Traffic and Wireless

Multimedia traffic has different characteristics than traditional internet traffic such as web browsing. Web browsing is characterized by multiple short bursts of data (as successive transactions between a web browser and a server generate and fulfill requests for data). Performance is largely dictated by achieving low latency and high burst data rates while guaranteeing that data is delivered error free. Since high sustained data rates are not necessary, it is feasible to share a channel with high capacity among multiple users. File transfer is another common internet traffic type; the emphasis here is on high average rate and error free transmission, but with allowable significant variation in instantaneous rate.

Real-time multimedia traffic makes two fundamentally different tradeoffs: a need for sustained data rates (at a rate that depends on the application, such as voice, video, music, etc.) and some tolerance for errors [1]. Latency is important if the multimedia is part of an interactive communication and less relevant if not (for example, one way streaming of a television program). Non real-time multimedia traffic has similar needs to those of large file transfers, albeit likely with some higher tolerance for errors than, say, a file of financial information. Of course, error correction can be handled at higher levels of the protocol stack. Quality of

Service (QoS) is a term used to connote the set of requirements of an application or the state of a network in terms of meeting those requirements. As we've seen, for multimedia these tend to focus on attributes of data rate (or "bandwidth"), rate variation, latency, and error rate.

Different physical layer networks face different challenges in meeting particular QoS expectations. Wireless networks have two particular characteristics of interest: a shared physical channel and a propensity for variations in error performance (caused, for example, by fading during relative motion of transmitter and receiver) [2].

Traditional cellular telephone networks solve the QoS problem for interactive voice by emulating the traditional circuit switched infrastructure of the public switched telephone network. Calls are blocked unless capacity can be dedicated to the call and the channel can be expected to have a low enough error rate to sustain intelligibility; what errors do occur are mitigated through error concealment.

Recent wireless networks, though, are much more likely to select a packet switched infrastructure rather than circuit switched, with the goal of efficiently handling non media applications. Indeed, this trend is evident in almost all networks. QoS for voice calls – and for other multimedia traffic – then becomes a configuration and management choice for the network rather than an intrinsic characteristic of the network design. These modern wireless networks are frequently designed first via natural extensions of computer local area network principles rather than as extensions of PSTN principles; the QoS needs of multimedia are either presumed solved adequately by abundant over provisioning or through explicit QoS semantic overlays on networks. In either case, though, we tend to exacerbate a fundamental characteristic of the economics of wireless internet access: the expectation and reliance on shared use.

3. Wireless Internet Access Economics and Oversubscription

Internet access services typically advertise and compete on the basis of data rate. For example, a DSL offering current as of the date of this paper advertised "up to 1.5 Mbps download/896 kbps upload" [3] while a competing cable company offer advertised "up to an unbelievable 6 Megs" but later caveats this with "Actual speeds may vary and are not guaranteed. Many factors affect speed." [4] One of the most important factors that affect speed is oversubscription.

Oversubscription, also known as statistical multiplexing, is really a statement of the relationship between a marketing or advertising contention about the rate that a customer could enjoy (e.g., 1.5 Mbps) and the amount of capacity available on the network if all users are actively attempting to secure the advertised rate. For example, if a network has a total capacity of 15 Mbps and is shared among 100 users all of who are told that they are purchasing a 1.5 Mbps service, the oversubscription ratio is 10 – the ratio between the 150 Mbps that would be required to serve all users simultaneously and the 15 Mbps that is actually available. Oversubscription ratios are a design choice to be made in any part of the network that is shared. In wireless and cable modems the last mile infrastructure is shared, though not in DSL. However, all networks share capacity in internet switching, routing, backhaul, and transport.

At first glance, oversubscription sounds like service provider fraud, but it is not for both legal and practical reasons. Legally, service providers use the terminology "up to" in describing speeds as well as other qualifications as we saw earlier in this section. Practically, as we indicated in the previous section, web surfing performance is largely determined by being able to support short high speed bursts. Providing request for bursts are not often coincident, it may well be true that a user on a 10x (or more) oversubscribed network sees the advertised rate (or even a better rate) almost anytime that it is important to the user experience.

Oversubscription ratio choices are usually considered proprietary information by service providers. This leaves end users to discipline service providers through a qualitative assessment of service quality and then expressing their preferences through purchase decisions. Of course, when there are only one or two broadband options in the market, expressing a preference can be somewhat limited. Because of the lock-in effects around service decisions as well as variability and noisiness in qualitative assessment, we can expect this effect to be present but slow and fickle. In spite of a lack of public information, industry trade publications suggest that service providers select oversubscription ratios ranging from 10 to 100 [5,6,7]. Wireless service providers, with both shared physical last mile links and shared transport and backhaul, engage in oversubscription.

The extent to which a particular physical network type is cost sensitive to oversubscription depends on the relative share of network cost that is devoted to shared infrastructure versus dedicated (per customer) infrastructure. The component of the network that is devoted to transport and backhaul of internet traffic tends both to be more common among different physical network types and a smaller fraction of total cost than the portion dedicated to last mile link access.

Consequently, most variation in sensitivity to oversubscription will arise from the last mile link.

DSL links do not share the last mile link; a dedicated copper twisted pair is used to carry traffic to each end user, although there is sharing in the DSL access multiplexer (DSLAM) at the telephone company's central office. Both cable systems and wireless internet access systems are more thoroughly shared. Cable systems share the bandwidth dedicated to cable modem use among a number of end users over a portion of the electromagnetic bandwidth on coax cables that reach all the end users. Similarly, wireless systems share capacity in the electromagnetic spectrum – either licensed or unlicensed – among users. Consequently, we can expect oversubscription to be a more important economic driver for cable and wireless networks than for DSL networks.

Further, there are subtle but important differences between cable and wireless networks. Both cable and wireless networks increase per user bandwidth (with increasing demand or increasing numbers of subscribes in a location) by geographic splitting: the subdivision of a shared area into smaller units that are independently shared. In cable plant this is called node splitting; in wireless plant it is called cell splitting. Cable node splitting involves adding additional electronics to a node or placing aggregation nodes deeper in the cable plant (closer to end users); importantly, though, these nodes are still located within the plant that the cable company already owns and maintains. Wireless cell splitting may involve the more expensive task of finding, renting or purchasing, and equipping suitable additional cell sites in real property not owned (or net yet owned) by the operator (new or existing towers, new building placements) as well as adding backhaul capacity to these locations. Moreover, a key piece of shared infrastructure, the wireless base station, is the subject of intentional cost shifting using advanced radio technologies – adding cost in the RF and signal processing portions of the base station in return for cheaper subscriber terminals. This is an eminently sensible choice in terms of overall system economics, but has the side effect of increasing the sensitivity of wireless system economics to oversubscription choices relative to those of other technologies.

The overall impact of these physical differences is that wireless network economics are likely to be more sensitive to oversubscription choices than either DSL or cable networks.

4. Oversubscription and Multimedia QoS

If wireless internet access network economics tend to be more sensitive to oversubscription ratios than other network types, it follows that they will be disproportionably impacted by changes in usage patterns that affect the relationship between oversubscription ratios and user perceptions of service.

In Section 2, we noted that multimedia traffic has differing QoS requirements. In particular, real-time multimedia traffic is (usually) not bursty but involves a sustained and steady demand for capacity. Non real-time multimedia traffic is similar to file transfers in that it benefits from high average capacity. However, multimedia files tend to be large relative to many other file types, so we can consider non-real time multimedia to represent the special case of large file transfers.

Unfortunately, both the case of a steady demand for capacity or the case of multiple large file transfers work directly against the acceptability of high oversubscription ratios for end users. In our 10x oversubscription example from section 3, if any ten of the 100 users attempt to access their "up to 1.5 Mbps" by streaming 1.5 Mbps video streams – rather than doing bursty web surfing – the remaining 90 customers are out of luck (or, more typically, all subscribers will see tangible service degradation). Similarly, if many users decide to simultaneously download as files whatever audio-video content they find compelling on their hard drive based computers or portable devices, all will quickly see much lower than the "up to" rates that were advertised a the network capacity saturates with sustained continuous data transfer sessions.

Hence, a shift towards increasing multimedia traffic as a fraction of total traffic on internet access networks will increase pressure on service providers to lower oversubscription ratios, and wireless network providers will be disproportionally affected relative to cable or DSL providers. To some extent the impact can be managed by explicit QoS configuration choices, but the overall impact is unavoidable.

5. Net Neutrality

In simple terms, network neutrality advocates contend that consumers should be able to access the content and applications of their choice without network carriers discriminating among these choices[1]. For example, under network neutrality a consumer would be able to use her Vonage phone service without the carrier blocking or otherwise degrading that service. However, there are a number of situations where a carrier might justify discriminating against certain content and applications. The obvious one would be whether the content was legal. Another would be based on network

[1] It should be pointed out that the concept of network neutrality is not new. It has its origins in various prior regulatory concepts of common carriage.

management requirements; in other words, the carrier claims that it is necessary to limit traffic flowing across its network in order to maintain certain performance expectations. Still another would be based on the need to limit access for network security reasons. This might include the need to prevent malicious reconnaissance such as vulnerability scanning or block the delivery of packets that might contain malware (such as computer viruses and worms). An arguably more contentious justification could be based on the claim that the content or application provider should compensate the carrier for the use of its network. Indeed, AT&T CEO Ed Whitacre was recently quoted as saying, "What [Google, Vonage, and others] would like to do is to use my pipes free. But I ain't going to let them do that." [8] Similarly, the CEO of Verizon, Ivan Seidenberg, stated that broadband application and content providers should "share the cost" of operating broadband networks. [9] In terms of wireless network neutrality, in November 2005 House Committee meeting, a Verizon Wireless executive stated that network neutrality should not apply to wireless carriers and indicated that carriers should be allowed to block traffic as they thought appropriate. [11]

In the policy statement Federal Communications Commission (FCC) 05-151,[2] the Federal Communications Commission stated four principles directed at ensuring that "broadband networks are widely deployed, open, affordable, and accessible to all consumers". [10] They include:

- To encourage broadband deployment and preserve and promote the open and interconnected nature of the public Internet, consumers are entitled to access the lawful Internet content of their choice.
- To encourage broadband deployment and preserve and promote the open and interconnected nature of the public Internet, consumers are entitled to run applications and use services of their choice, subject to the needs of law enforcement.
- To encourage broadband deployment and preserve and promote the open and interconnected nature of the public Internet, consumers are entitled to connect their choice of legal devices that do not harm the network.
- To encourage broadband deployment and preserve and promote the open and interconnected nature of the public Internet, consumers are entitled to competition among network providers, application and service providers, and content providers.

They also stated that they have "jurisdiction necessary to ensure that providers of telecommunications for Internet access or Internet Protocol-enabled (IP-enabled) services are operated in a neutral manner". [10] In other words, if they want to enforce network neutrality they have the authority. However, a footnote in the statement indicates that the principles are subject to "reasonable network management". This of course raises the question as to what constitutes reasonable network management; the concern here being that network management becomes a justification for selective content blocking and other similar measures.

In March of 2005, the FCC fined Madison River Communications (a rural phone company) for blocking connections to Internet phone providers (such as Vonage). As a result of FCC pressure, Madison River also agreed not to block such calls in the future. In this case, while several of the aforementioned justifications for blocking content were provided, the FCC found none of these as acceptable. [8] This action together with the policy statement suggests that the FCC does view network neutrality as a socially desirable goal.

This situation of network neutrality creates a number of interesting tensions. At the heart of this issue is the reasonable desire of the carriers to pass on the cost of network, but this is countered with the concern that the carriers will exhibit anticompetitive behavior and thereby erode much of what has made the internet successful. Telecommunications regulators have decided to move away from the heavily regulated model to embrace a model based on market discipline. However, with (at best) only two broadband players in most markets, it is difficult to belief that competition will produce the desired effect. We described earlier that the Internet is shifting away from the best-effort model where all bits are essentially treated equally to a model where certain bits can be given priority on the network; in other words, be provided differentiated service. As network carriers move to this differentiated service model, they will have the ability to decide whose bits are provided with priority treatment. This decision will likely be based on favoring their own bits or bits from carriers who pay for priority treatment. The concern is that those who don't pay could be relegated to a service level that doesn't support their application demands. What makes this a big concern is that the innovation of the internet has occurred not by the carriers but by the application providers.

In thinking about network neutrality within the wireless context, there exist an interesting history that might influence the thinking of both policy makers and consumers. Our common view of wireless is in terms

[2] It should be mentioned that these principles echo the earlier "Four Freedoms" described by Michael Powell.

of cellular telephony and WiFi (and likely more the former than the later). The cellular systems in the U.S. have a long history of using proprietary equipment to maintain control of equipment used on their networks. They also have a long history of embracing closed architectural choices and implementing differentiated pricing models for service selection (for example, paying extra to access the Internet on your cellular phone). In this sense, it should not be surprising to see these carriers taking steps that would violate some aspects of network neutrality. However, users have recently come to think of wireless to include WiFi and most implementations of WiFi (keeping with the traditional Internet model) do not violate network neutrality. As more handsets begin supporting both cellular and WiFi services, it will be interesting to see how wireless carriers respond to users that might make use of their handsets in ways that undermine potential profits for the wireless providers. It is not hard to imagine that a carrier would attempt to block handover to a "free" network or restrict the handover to a WiFi service that partners (i.e., pays) the carrier.

The issue of wireless network neutrality has arisen in a separate policy debate, namely municipal WiFi.[3] In a Request For Proposal (RFP) concerning the development of a citywide WiFi network put out by the city of San Francisco, reply comments contained language directly targeting issues of network neutrality. Quoting from the list of "demands", replies included (but were not limited to the following): [12]

- Open Access and Network Neutrality
- Assurances that free services don't lag substantially behind premium services, possibly requiring free services have at least one third the bandwidth of the premium service.
- Free service that is robust enough for general web use, email and messaging, and VOIP (in the initial rollout)

What makes this interesting is the explicit public call for network neutrality within a wireless network. Also of interest is the explicit demands for measurable bandwidth characteristics and support for realtime services such as VoIP. This shifts the network neutrality debate into defined and measurable expectations.

[3] The debate over municipal WiFi concerns whether or not a municipality should be allowed to build and/or offer a WiFi service. The carriers have opposed this concept and have contended (among other things) that municipalities are not capable of maintaining such networks. Their efforts to stop the build out of municipal WiFi have mostly focused on the state legislatures and they have been successful at getting a number of states to pass such legislation. The matter is now being considered at the Federal level.

6. The Interaction of Net Neutrality, Wireless, Multimedia, and QoS

At the end of Section 4, we noted that wireless service providers building new access networks around traditional shared access models that are effective for internet access like browsing will find themselves disproportionally affected relative to their DSL and cable competitors by a shift towards more multimedia traffic on such networks. Such a shift would be a consequence of the use of Voice over IP (VoIP) as an increasing or even a sole method of providing telephony applications and increasing streaming and downloading of music and video content on wireless networks. We also saw in Section 5 that a primary caveat to the principle of net neutrality is the need for "reasonable network management."

Service providers can make a plausible case that avoiding material degradation in the end user experience as the traffic mix shifts is an eminently practical form of "reasonable network management" and that either the service provider needs to manage the actual mix (by, for instance, restricting the amount of multimedia traffic allowed on the network or how it is treated through explicit QoS mechanisms) or by charging end users, application providers, or both at different rates depending on some measure of usage. At the same time, there will be an abiding concern that service providers might consider other reasons to stray from network neutrality, such as discriminating in favor of forms of traffic in which the service provider has a commercial interest. It will be challenging to separate these different rationales for deviation from network neutrality and an important subject of technical, economic, and regulatory analysis over the coming years. And, as we have seen, wireless internet access service providers are likely to be at the sharp point of this debate as they will feel most quickly and strongly the effects of shifts towards multimedia traffic.

7. Conclusion

We have argued here that a shift towards more multimedia traffic (such as VoIP, audio, and video) on internet access networks tends to have disproportionate impact on the economics of wireless network internet access providers compared to cable or DSL providers. This means that wireless providers are likely to feel most strongly the need to deviate from network neutrality – a policy that the FCC and others have argued is socially desirable - for reasons of "reasonable network management." The implication is that the

coming debate about how to separate reasonable exceptions to network neutrality from ones that are deemed anti-competitive or not in the public interest will be particularly sharp in the case of wireless providers.

As the policy debate and analysis around network neutrality continues, we can expect to use wireless access as an important test case for the impact of different policy approaches. Moreover, we will find net neutrality and other policy goals – such as the viability of wireless internet access as the so-called "third broadband pipe" in creating a competitive environment for broadband access – to become intermingled. Hence we see the evaluation of wireless network economics as a critical and ongoing component of this important policy debate.

8. References

[1] Wang, Zheng. *Internet QoS: Architectures and Mechanisms for Quality of Service*. Morgan Kauffman, 2001.

[2] Rappaport, Theodore. *Wireless Communications: Principles and Practice*, 2nd ed. Prentice-Hall, 2001.

[3] Qwest, "Residential: Internet/DSL," at https://iot.qwest.com/iot/internet/DslPickService.do?npa=303&nxx=413&sfx=1121&isp=&offerId=&prodDescCode=, visited 1/24/06.

[4] Comcast, "Special Offer" http://www.comcast.com/Benefits/CHSIDetails/Slot3PageOne.asp?LinkID=366, visited 1/24/06.

[5] Mitchell, Tom, "Avoiding the Pitfalls of Oversubscription in DSL Networks," *xchange*, at http://www.x-changemag.com/articles/041feat1.html, visited 1/24/06.

[6] Newman, Stagg. "Broadband Access Platforms," McKinsey and Company, April 14, 2005, at http://www.fcc.gov/oet/tac/april26-02-docs/BB-Access-Tech.pdf. visited 1/24/06.

[7] Earthlink Enterprise T1/T3 – Frequently Accessed Questions, 2005, at http://www.earthlink.net/biz/highspeed/enterprise/pdfs/Enterprise_FAQ.pdf, visited 1/24/06.

[8] At Stake: The Net As We Know It, Business Week Online, http://www.businessweek.com/technology/content/dec2005/tc20051215_141991.htm, posted December 15, 2005, visited January 24, 2006.

[9] Verizon Says Google, Microsoft Should Pay For Internet Apps, InformationWeek, http://www.informationweek.com/story/showArticle.jhtml?articleID=175801854, Posted Jan. 5, 2006, Visited January 24, 2006.

[10] FCC 05-151, Policy Statement, Federal Communications Commission, September 23, 2005.

[11] Verizon Wireless: Scrap Network Neutrality, http://techdirt.com/articles/20051110/1313223_F.shtml, Posted November 10th, 2005, Visited January 24, 2006.

[12] Webpage: The Tech of TechConnect, http://leftinsf.com/blog/index.php/archives/525, posted, December 21, 2005, visited January 24, 2006.

Spectrum Sharing Using Cognitive Radio Technology

Paul Greenis
Adapt4
Phone: (321) 259-5009
Email: paul@adapt4.com

Abstract--*Cognitive Radios (CRs) are a classification of software-defined radios (SDRs) that include complex, embedded software that 'harvests' wireless bandwidth. This means that the CR monitors a defined band of spectrum and determines which portions are in use at any instant in time. Portions that are not used are candidates for use by the CR. It is the method used for the harvesting that is of particular interest to regulators and spectrum owners. CRs must take every precaution to conduct harvesting on a 'non-interfering' basis. The presentation will discuss the critical aspects of sharing this unused spectrum while minimizing the probability of interfering with the primary user or licensee. Several concerns must be addressed in the design of the CR to ensure all precautions are taken. This includes such things as transmit power levels, waveform selection, frequency hopping and use detection sensitivity.*

In addition to optimizing the CR spectrum access techniques, several usage detection challenges must be overcome. There are several conditions which may exist that could cause 'undetected' interference with the primary users. These include 'hidden transmitters' and 'silent receivers' The presentation will address what methods can be taken to overcome these challenges. Additional capabilities of the SDR will be discussed which allow it to also address interoperability within a wide range of users, such as public safety organizations. It offers a powerful platform to support waveforms and access techniques necessary to communicate with existing, legacy radios as well as networking features which support the dynamic creation of user-defined communities. These features also make it a 'good fit' for military and Homeland Security applications.

Public Safety Environmental Noise Challenges for Land Mobile Radio Vocoders

D.J. Atkinson
Institute for Telecommunication Sciences
Phone – 303.497.5281
FAX – 303.497.5969
E-Mail – dj@its.bldrdoc.gov

Abstract

This paper investigates the effect on vocoders of background noise that exists in environments where public safety officials and first responders need to communicate. Four different environments were examined: police cruiser, fire engine, rescue boat, and rescue helicopter. During the examination, noise levels were measured, as well as speech levels of actual practitioners working in that environment. Based on those results a controlled laboratory experiment was conducted to determine the effectiveness of vocoders used in communications equipment subjected to those environments. The experimental conclusion is that while digital vocoders do not perform as well as analog transmission in the presence of high levels of background noise, moderate amounts of noise cancellation prior to the signal injection into the vocoder can somewhat mitigate those detrimental effects.

1. Introduction

Public Safety agencies continue to have insufficient spectrum to adequately meet their communications requirements, particularly in population centers. One facet to solving this problem is to increase the number of channels in the currently allocated spectrum. Currently in process is a migration from 25 kHz voice channels to 12.5 kHz voice channels across public safety agencies. One result of using traditional analog modulation in a narrowband channel is that the analog voice quality suffers significantly. Therefore digital voice coder/decoders (vocoders) are necessary to effectively accomplish the transition.

The Project 25 effort, a partnership between the Association of Public-Safety Communications Officials (APCO), the National Association of State Telecommunications Directors (NASTD), the Federal Partnership for Interoperable Communications (FPIC), and the Telecommunications Industry Association (TIA), has produced a suite of standards for a digital radio communication system that utilizes a low-bit-rate digital vocoder to implement these narrowband voice channels. When the vocoder was initially adopted in the early 1990s, there was a significant amount of testing done to ensure that the public safety community was getting the best audio quality available. However, one aspect that was overlooked in the investigation was how the vocoder would behave in the extremely high noise level environments in which public safety practitioners must operate. Consequently, there have been many practitioner complaints about voice quality levels on digital communication systems, particularly in environments with high background noise levels.

Currently, Project 25 is considering the effects of further narrowbanding to 6.25 kHz-equivalent channels. As part of their investigation into the continued channel squeeze, the Project 25 Steering Committee, in conjunction with the FPIC initiated an investigation that would provide scientific results showing the behavior of vocoders in the presence of high levels of background noise. Included in this study were the original Project 25 vocoder (7200 bits per second), a compatible vocoder at the same bit rate that includes enhancements developed in the years since the initial adoption, and a half-rate vocoder (3600 bits per second) in the same technology family.

This paper provides an overview of the investigation, including the noise measurement and sampling, and the design and conduct of the subjective test. Finally, a summary of the subjective testing results is presented.

2. Acknowledgements

The author of this paper wishes to express thanks to the following, whose assistance was essential to the success of the experiment:

- The City of Mesa, Arizona, and the US Coast Guard for providing opportunities to record high-quality public safety environmental noise.
- The Federal Partnership for Interoperable Communications (FPIC), sponsored by the DHS Office of the CIO.
- DHS Office of Interoperability and Compatibility's SAFECOM program, which provided travel arrangements for the listening subjects.
- The following communities/agencies for adjusting schedules and allowing their public safety personnel to travel to Boulder to participate in the experiment: City of Mesa, Arizona; City of Phoenix, Arizona; County of Los Angeles, California; City of Roseville, California; Yolo County, California; California Highway Patrol; City of Gurnee, Illinois; North Carolina State Highway Patrol; City of Alexandria, Virginia; Four Mile Rural Volunteer Fire Department, Colorado; US Postal Inspection Service.

3. Environmental Noise Sampling

To adequately determine signal-to-noise ratios (SNRs) at the microphone of a public safety practitioner's radio it was necessary to record and take sound pressure level (SPL) measurements of noise and speech+noise in several common environmental situations.

3.1 Noise Sampling

For each noise condition, a 48 kHz[1] sampling 16 bit-per-sample recording was made with two Beyerdynamic MCD 100 digital microphones[2] and a portable digital audio tape (DAT) recorder. In addition, SPL measurements were made.

For vehicle interior sampling, the microphones were separated by six inches and placed at ear level in a position equivalent to where a center passenger would sit in the front of the vehicle. The Beyerdynamic MCD 100 digital microphones have a cardioid pattern. The microphones were positioned such that the cardioid pattern was 60 degrees offset from straight ahead and crossing the center plane of the vehicle (i.e., the left microphone records sounds from the right half of the vehicle and vice versa). Figure 1 illustrates the dual cardioid pickup pattern achieved by this microphone placement, which provided excellent stereo separation while still providing good coverage from front-center, where a significant amount of noise emanates. The back-center is assumed to be the lowest noise point, and therefore it is reasonable to minimize sensitivity at this point in the recording field.

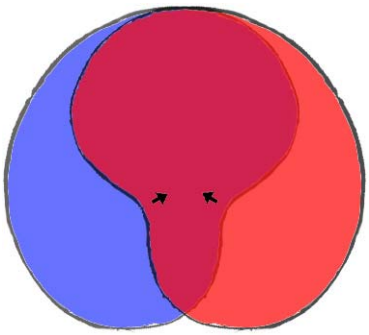

Figure 1. Dual cardioid pickup pattern.

For vehicle exterior sampling, the microphones were placed at 63 inches above ground (average ear height of a person of average height) and 5 feet from the passenger door of the vehicle in the same configuration as just described for interior sampling and shown in Figure 1, with the change that the pickup pattern was focused toward the vehicle.

SPL measurements were conducted using a Galaxy Audio CM-150 Sound Level Meter in both A-weighted and C-weighted[3] modes. For each sampling condition, measurements were made in the "forward" directions and at 60-degree increments throughout the horizontal plane. In the forward position, average, peak and minimum SPL were

[1] Sampling at 48 kHz provides audio frequency response equivalent to (or better than) that of a compact disc and provides an integral divisor when downsampling to 8 kHz, the sampling rate used at the input to the voice coder.

[2] Certain commercial equipment, materials, and/or programs are identified in this report to specify adequately the experimental procedure. In no case does such identification imply recommendation or endorsement by the National Telecommunications and Information Administration, nor does it imply that the program or equipment identified is necessarily the best available for this application.

[3] A-weighted SPL measurements correlate with loudness for softer signals, while the broader band C-weighted SPL measurements correlate with loudness for louder signals.

measured. Only average SPL was measured in the other positions. Recording conditions and measured average SPL are summarized in Table 1. SPLs presented in the table are average values for the forward case. Accuracy of the instrumentation is ±1.5 dB.

Table 1. Summary of Noise Conditions and Levels.

Condition	Average Sound Pressure Level (SPL)	
	A wt.	C wt.
Fire – Idle – Interior	70 dB	86 dB
Fire – Fast Idle – Interior	84 dB	94 dB
Fire – Idle – Exterior	79 dB	88 dB
Fire – Fast Idle – Exterior	87 dB	98 dB
Fire – 30 MPH – No Siren	82 dB	94 dB
Fire – 30 MPH – Siren	88 dB	96 dB
Fire – 60 MPH – No Siren	86 dB	110 dB
Fire – 60 MPH – Siren	87 dB	110 dB
Police – 30 MPH – No Siren	62 dB	86 dB
Police – 30 MPH – Siren	75 dB	86 dB
Police – 60 MPH – No Siren	67 dB	91 dB
Police – 60 MPH – Siren	70 dB	88 dB
Helicopter – Exterior	105 dB	107 dB
Helicopter – Interior Idle	95 dB	107 dB
Helicopter – Interior Flight	96 dB	107 dB
Boat – 4.8 knots	69 dB	89 dB
Boat – 30 knots	88 dB	92 dB

Because a subjective test evaluating all of these noise environments would be unreasonably large, a subset of these environments was chosen for the subjective experiment. The environments chosen were:

- Fire Truck – 30 MPH – Siren
- Fire Truck – 60 MPH – No Siren
- Boat – 30 knots
- Helicopter – Interior Flight
- Police – 60 MPH – Siren

Figure 2 provides the power spectral density (PSD) plots for these noise samples, and shows that for most public safety noise environments, the majority of the noise power is very low frequency. The exception to this is the noise on an outboard boat, which has a peak power at approximately 200 Hz.

3.2 Speech+Noise Sampling

In addition to the noise recordings, a number of speech+noise recordings were made at 48 kHz sampling, 16 bits per sample. These recordings were made with three different microphones: a Beyerdynamic MCD 100 digital microphone, a Motorola HMN1061A ("Spectra") microphone, and a David Clark M-1/DC amplified dynamic noise canceling microphone. In addition, SPL measurements were made.

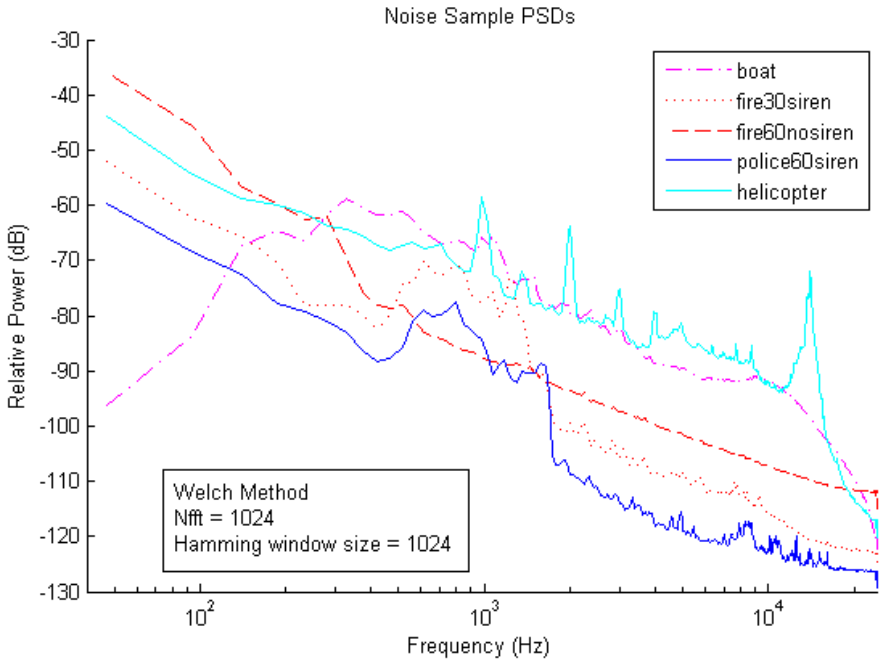

Figure 2. Power spectral density plots for noise conditions used in subjective vocoder experiment.

For speech+noise recording and measurements, the microphone was hand-held at the mouth reference point.[4] For interior samples, the talker was positioned in the front passenger/navigator/copilot seat facing the front of the vehicle. For exterior samples, the talker was positioned five feet from the passenger door of the vehicle and facing toward the vehicle. The recordings and measurements were used to identify SNRs in environments where public safety radio equipment is used. Table 2 summarizes the C-weighted SNRs observed with the CM-150 for the conditions used in the subjective tests.

Table 2. Summary of SNRs Observed in the Field.

Condition	SNR
Fire – 30 MPH – Siren	3 dB
Fire – 60 MPH – No Siren	-11 dB
Police – 60 MPH – Siren	9 dB
Helicopter – Interior Flight	-5 dB
Boat – 30 knots	0 dB

Because three of the conditions were considered extreme by the TIA TR8 Vocoder Task Group, they requested that additional conditions be added with a 6 dB gain in SNR, to mimic the use of a noise-canceling microphone. The added conditions were: Fire – 60 MPH – No Siren @ -5 dB SNR, Helicopter – Interior Flight @ 1 dB SNR, and Boat – 30 knots @ 6 dB SNR.

4. Subjective Experiment

4.1 Experiment Design

Design of a subjective speech quality assessment experiment is an exacting process. Several Recommendations from the International Telecommunication Union (ITU) address factors in the design of subjective experiments. The two primary recommendations that served as the guide for this experiment were ITU-T Recommendation P.800 [3] and ITU-T Recommendation P.830 [4].

Based on those Recommendations, a subjective voice quality experiment typically proceeds as follows. A test subject, or listener, is seated in a sound-attenuated chamber. Inside the chamber, the subject uses headphones to first set the volume at his/her preferred level. The subject then listens to several 5- to 7-second long pairs of sentences, rating the quality of each pair of sentences in turn. The subject is asked to provide their opinion of the overall quality of the speech that they hear on a rating scale that uses the terms “Excellent”, “Good”, “Fair”, “Poor”, and “Bad”. Quality ratings are converted to numbers on a five-point scale where 1 represents a “Bad” rating and 5 represents an “Excellent” rating. These scores are then averaged across listeners, talkers, and sentence pairs to compute an MOS.

The listening material was developed based on the following factors:

- Talkers: 4 male, 4 female
- Sentence pairs: 5 per talker
- Input level: -28 dB_{OV}[5] [5] [6]
- Vocoders: P25 Full Rate (FR)[7], P25 proposed Enhanced Full Rate (EFR), P25 proposed Enhanced Half Rate (EHR), and 16-bit linear PCM as a reference
- Background noise conditions: Clean (no background noise), Fire Truck (60 MPH, No Siren – 2 noise levels), Fire Truck (30 MPH, Siren), Police Cruiser (60 MPH, Siren), Helicopter (Cruising – 2 noise levels), Coast Guard Boat (30 Kts – 2 noise levels)
- Control conditions: modulated noise reference unit (MNRU) [8] at -5, 5, 15, and 25 dBQ.[6]

In addition to the use of realistic background noise, another novel factor in this experiment is the use of actual public safety practitioners as listeners. This experiment used 27 listeners that represented law enforcement, emergency medical services, fire, dispatch, and federal users.

4.2 Experiment Development Issues

During the development of the experiment, two issues arose that required some additional consideration to produce acceptable results. These were proper mixing of the speech and noise, and signal levels output from the MNRU tool at low signal-to-distortion ratios.

4.2.1 Mixing Speech and Noise

Because the noise levels used in this experiment were so high, it was critical that the mixing of speech and noise be done as accurately as possible to achieve the same levels that were observed in the field in such a way that the speech portion of the signal was always at -28 dB_{OV} when it was input to the vocoder or MNRU. Factors addressed included 1) compensation for the weighting factors used in the field

[4] The mouth reference point is 25 mm in front of the lip plane as defined by ITU-T Recommendation P.51[2].

[5] dB relative to the overload point of a digital system. In this case, the digital system is 16 bits per sample.

[6] dB of signal to distortion ratio

measurements, and 2) downsampling from the 48 kHz sampled recordings to the 8 kHz sampled files that were input into the vocoder software and played to the listeners. Production of a scaled, mixed, 8 kHz sampled speech file is accomplished in three stages: 1) computation of speech gain factor G_S, 2) computation of noise gain factor G_N, and 3) creating the file to be presented to the coder. The process is shown in Figure 3.

G_S is a gain correction factor required to present the 8 kHz sampled speech to the vocoder at -28 dB_{OV}. G_S is computed by taking 48 kHz sampled speech, downsampling to 8 kHz [6], and computing the active signal level [5][6] (ASL1). Once ASL1 is known, G_S is computed according to:

$$G_S = -28 - ASL1$$

G_S is computed for each sentence pair to be passed through the codec. G_N is the gain factor required to achieve the noise level such that mixing (summing) the noise and the scaled speech signal will result in the SNR specified in this experiment. Those SNRs were developed using measurements taken using a C-weighted scale, which is the appropriate weighting for environments above 85 dB_{SPL}[9]. Therefore, the gain correction must take into account the C-weighting to ensure that the SNR presented to the speech coder represents the conditions measured in the field.

First, the active signal level of the C-weighted speech must be computed. This is done by applying G_S to the 48 kHz sampled speech, filtering the speech through a C-weighted filter[7] (Wc)[10], and computing the active signal level (ASL2) [5][6].

Next, the active signal level of the C-weighted noise must be computed. This is done by mixing (summing) the left and right channels of the stereo noise recording, filtering the noise through a C-weighted filter (Wc), and computing the active signal level (ASL3) [5][6].

Using the SNR specified for the particular condition, G_N can then be computed according to:

$$G_N = (ASL2 - ASL3) - SNR$$

G_N is computed for each speech file/noise file combination to be mixed.

[7] The frequency domain amplitude transfer function for the C-weighted filter is given by:

$$Rc(f) = \frac{12200^2 * f^2}{(f^2 + 20.6^2)(f^2 + 12200^2)}$$

where f is frequency in Hz.

Prior to mixing (summing) with the noise signal, the speech is scaled according to G_S. This ensures that the speech will be at -28 dB_{OV} after the speech signal is downsampled to 8 kHz.

Prior to mixing (summing) with the speech signal, the noise signal is scaled according to G_N. This ensures that the SNR of the mixed file will be correct according to the C-weighted measurements recorded in the field and specified in this experiment.

The scaled speech and scaled noise are mixed (summed), and downsampled to 8 kHz[6].

The final step in the process is the addition of a high-pass filter (HPF) prior to injection of the signal into the codec. This provides protection of the codec against the extremely high levels of low-frequency noise encountered in some of the environments being simulated. In addition, the filter provides a response that is representative of that provided by a microphone that might be used with a radio implementing the codec.

The HPF implemented here is a 2nd order IIR filter with F_c = 20.6 Hz and F_s = 8000 Hz. The Z-domain representation of this filter is given by:

$$Y(z) = 0.984 \cdot \frac{1 - 2z^{-1} + z^{-2}}{1 - 1.9679z^{-1} + 0.96816z^{-2}} X(z)$$

This filter was derived from the high-pass portion of the C-weighted filter used to compute the SNR.

4.2.2 MNRU Signal Level

During the development of the experiment, there was some concern that the high noise levels, relative to the -28 dB_{OV} speech signal, might cause clipping to occur in some speech files. During the process of ensuring that none of the mixed files contained clipped samples, clipping was observed in the output of the MNRU software tool at the -5 dBQ level. When noise is added to speech, and the noise level is similar to, or higher than, the speech level, the resulting combined signal will have a level that can be significantly higher than the original speech signal. The curve representing this effective gain factor is shown in Figure 4. At levels higher than 10 dBQ, the gain factor is negligible, less than 0.5 dB. At levels lower than 10 dBQ (5 and -5 dBQ for this experiment) the gain factor had to be compensated for by adjusting the level of the speech file prior to input into the software tool. The gain correction factors applied to the 5 and -5 dBQ conditions were -1.2 dB and -6.1 dB, respectively.

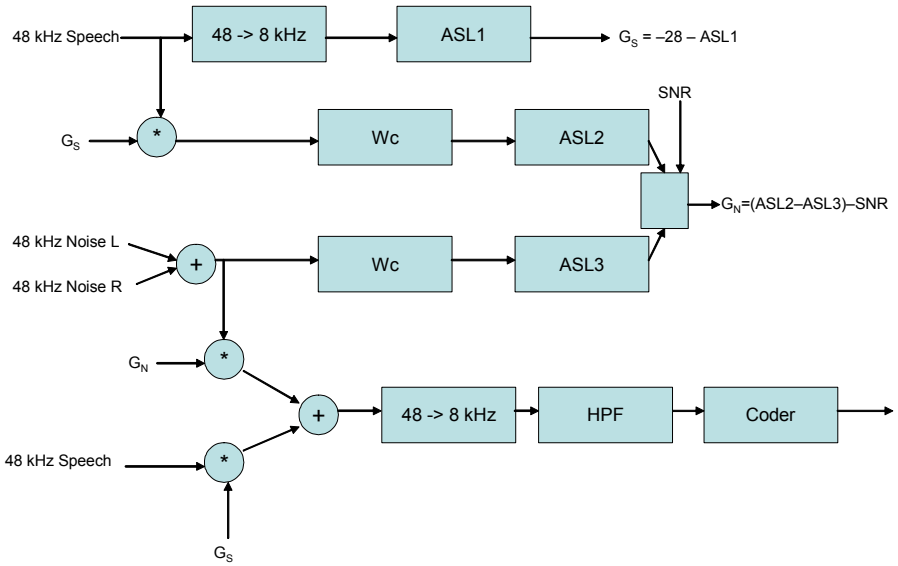

Figure 3. Process for producing a mixed speech plus noise file.

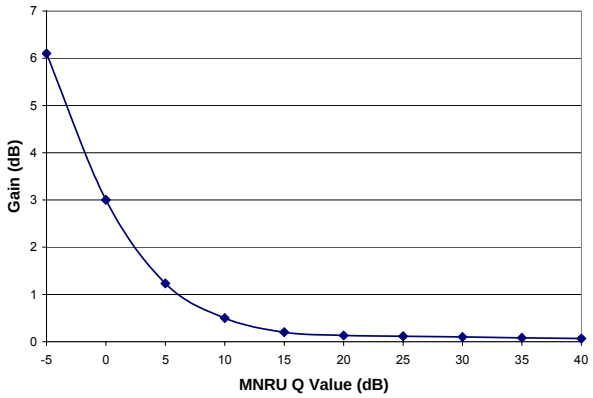

Figure 4. MNRU Gain by Q Value (dB)

5. Experiment Results

Figure 5 contains a bar chart of Mean Opinion Scores (MOSs) for the three vocoders under test and the reference conditions. The pattern in each bar repesents the background noise condition. Visual examination of the figure shows that there are differences between the reference conditions and the three vocoders, and that the two enhanced vocoders are better than the original full-rate vocoder.

Table 3 contains the results for all 40 conditions included in this experiment. MOS values vary from

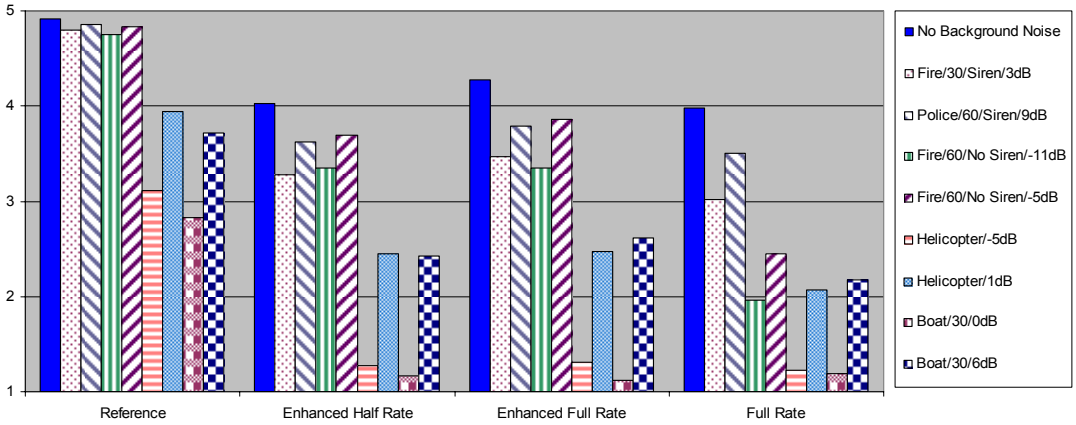

Figure 5. Selected Mean Opinion Scores.

Table 3. Mean Opinion Scores and Standard Deviations for All Conditions

Vocoder	Background Noise	MOS	Sdev
Linear PCM		4.913	0.082
25 dBQ MNRU		4.221	0.177
15 dBQ MNRU		3.279	0.162
5 dBQ MNRU		2.236	0.139
-5 dBQ MNRU		1.096	0.092
Enhanced Half Rate	Fire 30 Siren (3 dB)	3.284	0.221
Enhanced Full Rate	Fire 30 Siren (3 dB)	3.471	0.302
Full Rate	Fire 30 Siren (3 dB)	3.014	0.203
Reference	Fire 30 Siren (3 dB)	4.798	0.084
Enhanced Half Rate	Police 60 Siren (9 dB)	3.620	0.255
Enhanced Full Rate	Police 60 Siren (9 dB)	3.788	0.220
Full Rate	Police 60 Siren (9 dB)	3.500	0.127
Reference	Police 60 Siren (9 dB)	4.856	0.094
Enhanced Half Rate	Fire 60 No Siren (-11 dB)	3.346	0.276
Enhanced Full Rate	Fire 60 No Siren (-11 dB)	3.351	0.254
Full Rate	Fire 60 No Siren (-11 dB)	1.957	0.357
Reference	Fire 60 No Siren (-11 dB)	4.750	0.074
Enhanced Half Rate	Fire 60 No Siren (-5 dB)	3.697	0.291
Enhanced Full Rate	Fire 60 No Siren (-5 dB)	3.865	0.240
Full Rate	Fire 60 No Siren (-5 dB)	2.452	0.426
Reference	Fire 60 No Siren (-5 dB)	4.832	0.041
Enhanced Half Rate	Helicopter (-5 dB)	1.274	0.222
Enhanced Full Rate	Helicopter (-5 dB)	1.303	0.242
Full Rate	Helicopter (-5 dB)	1.231	0.190
Reference	Helicopter (-5 dB)	3.115	0.340
Enhanced Half Rate	Helicopter (1 dB)	2.452	0.256
Enhanced Full Rate	Helicopter (1 dB)	2.466	0.375
Full Rate	Helicopter (1 dB)	2.067	0.376
Reference	Helicopter (1 dB)	3.938	0.225
Enhanced Half Rate	Boat 30 knots (0 dB)	1.163	0.096
Enhanced Full Rate	Boat 30 knots (0 dB)	1.115	0.139
Full Rate	Boat 30 knots (0 dB)	1.188	0.108
Reference	Boat 30 knots (0 dB)	2.822	0.288
Enhanced Half Rate	Boat 30 knots (6 dB)	2.428	0.329
Enhanced Full Rate	Boat 30 knots (6 dB)	2.615	0.307
Full Rate	Boat 30 knots (6 dB)	2.178	0.354
Reference	Boat 30 knots (6 dB)	3.716	0.291
Enhanced Half Rate	No Background Noise	4.029	0.367
Enhanced Full Rate	No Background Noise	4.274	0.183
Full Rate	No Background Noise	3.976	0.279

low of 1.096 to a high of 4.913 (5.0 is the maximum possible score). Standard deviations vary from 0.082, which is typical for the extreme ends of the quality range (i.e., 1 and 5) to 0.376, which is slightly larger than desired for a subjective experiment. The latter might have been mitigated through the use of a larger subject pool.

5.1 Observations

Because the listeners were public safety practitioners, they were somewhat different from typical "naïve" listeners. Most listeners will take a break when it is offered – in this case after each of the four sessions. However, many of the practitioners wanted to complete the whole experiment without taking the offered break between sessions. Some of the responses included "I have to listen to this type of material for 12 hours a day, this is nothing," and "This is easier than work because there's only one person talking at a time."

6. Summary of Results

Table 4 summarizes the statistically significant differences among the vocoders and between the vocoders and the reference condition by showing which vocoder was better if there was a difference or "Equiv" if they were equivalent. It also indicates whether any of the vocoders produced a speech signal that would be at all useable, where useable is defined as a MOS greater than 2.0.

Significant observations in the table are:

- There were two conditions (helicopter @ -5 dB SNR and boat at 0 dB SNR) for which none of the vocoders produced a speech signal of acceptable quality. These conditions are the only ones where all three vocoders were statistically equivalent.
- For those two conditions mentioned in the previous bullet, an increase of 6 dB SNR (as might be achieved from a noise-canceling microphone) made a significant difference in the quality of speech produced by the two enhanced vocoders (at least 1 full MOS point increase).

Table 4. Summary of Statistically Significant Differences Among Vocoders Tested.

	Statistically Significant Difference From			Any vocoder
	FR to EFR	FR to EHR	EHR to EFR	>2.0 MOS?
No Background Noise	EFR	Equiv	EFR	Yes
Fire 30 Siren (3 dB)	EFR	EHR	Equiv	Yes
Police 60 Siren (9 dB)	EFR	Equiv	Equiv	Yes
Fire 60 No Siren (-11 dB)	EFR	EHR	Equiv	Yes
Fire 60 No Siren (-5 dB)	EFR	EHR	Equiv	Yes
Helicopter (-5 dB)	Equiv	Equiv	Equiv	No
Helicopter (1 dB)	EFR	EHR	Equiv	Yes
Boat 30 knots (0 dB)	Equiv	Equiv	Equiv	No
Boat 30 knots (6 dB)	EFR	EHR	Equiv	Yes

- For all conditions except those where no vocoder had a MOS > 2.0, the Enhanced Full Rate vocoder was significantly better than the Full Rate vocoder.
- For five of the seven conditions for which one or more vocoders produced acceptable quality, the Enhanced Half Rate vocoder was significantly better than the Full Rate vocoder.
- The Enhanced Full Rate vocoder was significantly better than the Enhanced Half Rate vocoder for only one condition: the reference condition (no background noise).

7. References

[1] ITU-T Recommendation G.711, "Pulse code modulation (PCM) of voice frequencies."

[2] ITU-T Recommendation P.51, "Artificial Mouth."

[3] ITU-T Recommendation P.800, "Methods for subjective determination of transmission quality."

[4] ITU-T Recommendation P.830, "Subjective performance assessment of telephone band and wide band digital codecs."

[5] ITU-T Recommendation P.56, "Objective measurement of active speech level."

[6] ITU-T Recommendation G.191, "Software tools for speech and audio coding standardization."

[7] ANSI/TIA/EIA 102.BABA, APCO Project 25 Vocoder Description, May 1998.

[8] ITU-T Recommendation P.810, "Modulated noise reference unit (MNRU)."

[9] ANSI S1.4, "Specification for Sound Level Meters," 2001.

[10] ANSI S1.42, "Design Response of Weighting Networks for Acoustical Measurements," 2001.

Radio Communications for Emergency Responders in Large Public Buildings: Comparing Analog and Digital Modulation*

Kate A. Remley [1], Marc Rütschlin, Dylan F. Williams,
Robert T. Johnk, Galen Koepke, Christopher L. Holloway
National Institute of Standards and Technology; 325 Broadway; Boulder, CO 80305
1. Corresponding author: ph. (303) 497 3652, email: remley@boulder.nist.gov

Mike Worrell, Phoenix Fire Department; Phoenix, AZ
Andy MacFarlane, Medaire, Inc.; Phoenix, AZ

To assess in-building radio coverage, in 2004 the City of Phoenix Fire Department carried out extensive testing of their radio systems. They deployed firefighters in standard configurations in a variety of buildings, and rated on a scale of 1 to 5 the audio quality of the received signals. To provide a link between the qualitative ratings and absolute field strength, NIST staff later carried out a series of measurements side-by-side with the Phoenix firefighters. The calibrated data from the NIST tests enables translation of the larger set of Phoenix Fire Department data into transferable values useful to industry, standards organizations, and other public-safety groups. We report here on a subset of these tests that compare analog and digitally modulated signals at 800 MHz.

1. Introduction

Reliable communication between emergency responders operating in hazardous situations is critical to both the safety of personnel and the success of their mission. Their radio communications equipment must be extremely reliable, and the communications functions provided must be predictable. To assess their in-building radio communications under field conditions, the Phoenix Fire Department (PFD) conducted extensive testing of new and existing radio systems deployed in configurations common to the department [1].

Testing focused on the ability of firefighters deployed in a building interior to be able to communicate with a command position on the exterior of the building. An analysis of fire-ground communications was performed for each building studied. All responses were based on the Phoenix Fire Department Standard Operating Procedures (SOPs) for the various National Fire Protection Association (NFPA) building types (see Appendix).

During the testing process, the National Institute of Standards and Technology (NIST) became aware of the Phoenix Fire Department testing. NIST developed measurement techniques to determine the absolute received signal strength at locations near PFD personnel during a set of tests. A team from NIST spent a week measuring radio signal levels. The NIST measurements were later calibrated in a post-processing step to find the electric field strength. Comparing these field strengths to the qualitative ratings provided by Phoenix Fire personnel allowed us to "translate" the larger set of Phoenix Fire Department tests into approximate quantitative values that are transferable to other locations and enables technical evaluation or comparison of different radio systems.

Among the tests carried out were those using point-to-point 700/800 MHz analog simplex and 700/800 MHz digital simplex. These tests provided an opportunity to compare radio reception when both digital and analog modulation were used. We report on these tests below.

2. The Phoenix Fire Department Tests

The Phoenix Fire Department tests were performed over an eight-week period. The tests were conducted in 30 buildings that consisted of the five different NFPA construction types (see Appendix). Approximately 1,500 talk paths were tested. The same test participants were used throughout the testing to provide test consistency in the grading process.

A command structure was developed for each response. Analyzing the command structure and determining the logical fire ground communications paths led to the development of the talk matrix. A test script was developed for each unique communications path identified in the talk matrix. Each communications path was categorized as either fire ground communications or wide-area communications, with the majority of the paths being fire ground. A typical layout is shown in Fig. 1.

Each building was pre-planned for the test session. Personnel were placed in the buildings to represent fire companies on an incident response. The personnel

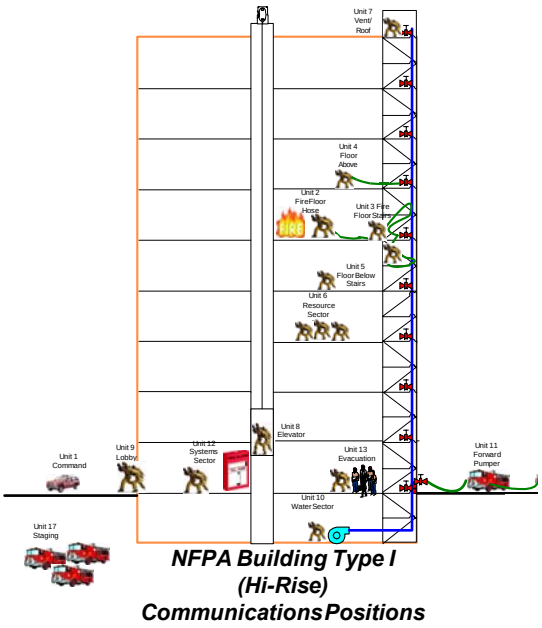

Figure 1: Typical firefighter deployment in a high-rise building. Radio links between all deployed positions were tested by PFD.

followed the test plan and graded each communication path. Each path was graded on a 1-5 scale with 1 representing poor communications and 5 being the best audio quality. Participants also determined whether the communications were useable on the fire ground on a pass/fail basis.

Upon completion of testing each day, the building's test results were entered into a spreadsheet containing the bidirectional grades for each communications path. Using the spreadsheet data, histograms were created for each building type and path category. NFPA Type 1 (concrete and steel) buildings were found to have the most variability in the histograms, and are also the buildings with the highest-risk environments, most complex interiors and largest operational structure. The histograms in the following are for fire-ground operations in Type 1 buildings.

Test results showed that the 700/800 MHz analog simplex channels provided clear, consistent communications in most test situations. This is shown in Fig. 2 by the high number of "5" ratings in the histogram. The test participants preferred this mode of operation over all others. Before testing commenced we expected that there would be differences in penetration capabilities between the 450 MHz and 700/800 MHz RF frequency bands. However, the penetration differences between the RF bands were negligible for the buildings we tested.

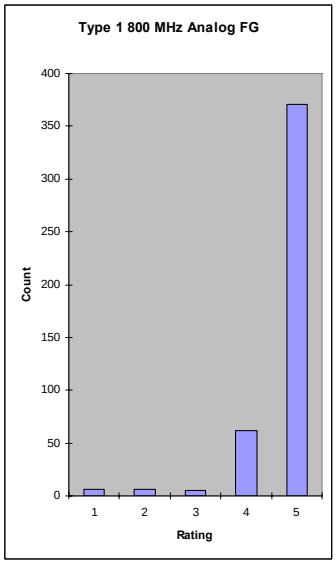

Figure 2: PFD test results: Analog simplex signals received in the local fireground (FG) over the course of all tests. A rating of 5 represents the clearest audio transmission.

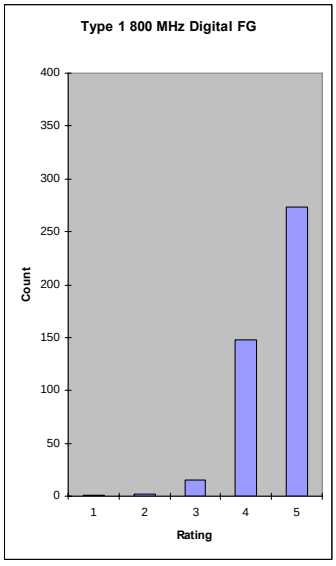

Figure 3: PFD Test Results: Digital simplex signals received in the local fireground (FG) over the course of all tests. A rating of 5 represents the clearest audio transmission.

The 700/800 MHz digital simplex channels (see Fig. 3) provided consistent communications. Test participants did note that the majority of transmissions had some level of digital distortion as reflected by the large number of "4" ratings and an increase in "3" ratings. The typical level of distortion did not render the communications unusable. However, users did encounter more situations where a repeat broadcast was needed to interpret the message.

Digital modulation did outperform analog modulation in many low signal-level situations. The digital mode would

Figure 4: NIST calibrated measurements were taken in close proximity to Phoenix Fire personnel using the system shown on the left side of the photo.

provide understandable communication when the same analog path would be scratchy and barely readable.

3. The NIST Tests

The goal of the NIST collaboration with the Phoenix Fire Department was to assign measured power and absolute electric field strength levels to the various qualitative ratings. While rating schemes such as the one described above are common, they suffer limitations inherent to "subjective" scales: they do not enable technical evaluation or comparison of different radio systems and it is difficult to compare ratings carried out by other groups in similar experiments.

The NIST measurement system consisted of a communications receiver, antenna system, and laptop computer. This system was developed to detect very weak signals in a cost effective way. The calibrations involved in these tests are described in [2,3]. The receiver system was used to collect data on the ground floor of an eight-story commercial building in Phoenix. The transmitter – a portable radio identical to those used by the Fire Department – was carried up and down a staircase on the opposite side of the building.

To ensure that the NIST experiments covered the entire range of voice quality ratings described in the Phoenix tests, we carried the transmitters to places with good and bad radio reception. At various locations in the building, an audio and silent test count were performed to allow nearly simultaneous PFD evaluation of the voice quality and NIST measurement of the power in the modulated carrier. The quality of the audio transmission was evaluated by a firefighter located at the listening station; both test counts were recorded using the communications

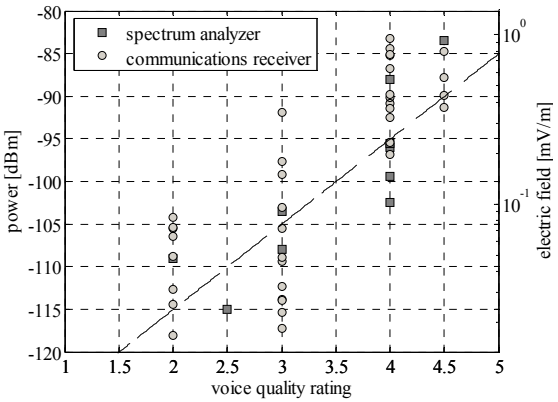

Figure 5: Measured average power and field strengths corresponding to voice quality evaluations for 860 MHz analog transmissions. The dashed line is a linear fit of all the data.

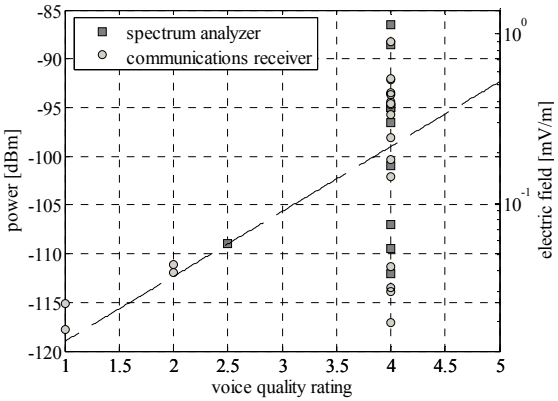

Figure 6: Measured average power and field strengths corresponding to voice quality evaluations for 860 MHz digital transmissions. The dashed line is a linear fit of all the data.

receiver system for later calibration and evaluation. The silent test counts were also monitored with a spectrum analyzer. This allowed us to have a backup set of measurements should one system fail, and also provided an independent verification of the communications receiver measurements. Among other tests carried out, analog and digital transmissions at 860 MHz were investigated and are the ones reported on here.

Scatter plots of the measured field strengths corresponding to voice quality evaluations are shown in Figures 5 and 6 for the 860 MHz analog and 860 MHz digital transmission respectively. The linear fit to the data in each graph may be used to estimate the field strength required for a particular quality of transmission. For example, in Fig. 5 all of the analog transmissions rated 3 correspond to received power levels below -90 dBm.

Unlike the analog case, the digital signal scatter plot in Fig. 6 shows that the majority of transmissions were given a voice quality level of 4, even for quite weak measured signal levels. This is the "digital shelf" effect noted by the PFD earlier. Note that all the quality ratings of 3 or worse were for signals with power levels of lower than -105 dBm.

The number of ratings of lower audio quality in the NIST tests is greater than that for the Phoenix Fire Department tests shown in Figs. 2 and 3. The NIST-led experiment was designed to intentionally cover the entire range of audio quality levels, whereas the Phoenix Fire tests reflected more realistic communications scenarios.

Factors such as multipath, the approximate one-meter difference in position between firefighters and the receiver system, repeatability of transmitter position, and the subjective nature of the audio quality assessments likely contribute to the spread of data for a given audio quality rating in Figs. 5 and 6. The graphs shown here are thus an estimate of the range of signal strengths that would result in a certain voice quality being received in the same room.

4. Summary

Tests that compared audio quality of received signals from analog and digitally modulated radio systems were part of an extensive study carried out by the Phoenix Fire Department. The purpose of this study was to assess effectiveness and deployment of various firefighter radio systems in the field. Results showed that both analog- and digitally-modulated systems work in a majority of situations in an NFPA Type 1 building (concrete and steel). However, digitally modulated signals were discernable at lower signal levels than those for their analog counterparts. When both types of signals could be received, the firefighters preferred the audio quality of the analog signals.

To investigate the link between the subjective ratings used in the Phoenix Fire Department tests and received signal strength levels, NIST carried out measurements of absolute field strength in an eight-story building in Phoenix in which poor signal transmission quality had been observed in previous tests. Audio quality ratings were provided by members of the Phoenix Fire Department who had participated in the earlier tests. Plotting the received signal levels vs. audio quality ratings provided a link between quantitative and the qualitative results. This link allows for comparison of the Phoenix Fire Department test data with data collected by other groups. Furthermore, having absolute signal strengths enables the data to be used in the technical evaluation of different radio systems, which may be of use to industry, standards organizations, and other public-safety groups.

Appendix—NFPA Building Types

Type 1: Fire-Resistive Construction
Reinforced concrete and structural steel

Type 2: Non-Combustible/Limited Combustible Construction.

a) Metal-Frame covered by metal exterior walls
b) Metal frame enclosed by concrete block, non-bearing exterior walls
c) Concrete block bearing walls supporting a metal roof

Type 3: Ordinary Construction/Brick and Joint Construction

Type 4: Heavy Timber Construction

Type 5: Wood Frame Construction

Acknowledgement

This work was funded in part by the Department of Justice Community Oriented Police Services (COPS) and the Department of Homeland Security through the NIST Office of Law Enforcement Standards.

References

[1] Mike Worrell and Andy MacFarlane, "Phoenix Fire Department Radio System Safety Project," *Final Report*, Oct. 2004. Report available at http://www.ci.phoenix.az.us/FIRE/radioreport.pdf Copyrighted material used by permission.

[2] M. Rütschlin, K. A. Remley, R. T. Johnk, D. F. Williams, G. Koepke, C. Holloway, A. MacFarlane, and M. Worrell, "Measurement of weak signals using a communications receiver system," *Proc. Intl. Symp. Advanced Radio Tech.*, Boulder, CO, March 2005, pp. 199-204.

[3] M. Rütschlin, K. A. Remley, D. F. Williams, R. T. Johnk, G. Koepke, D. Camell, M. Worrell, and A. MacFarlane, "Field-strength measurement of weak-modulated signals in buildings," *in process*.

RFID-Assisted Indoor Localization and Communication for First Responders

L. E. Miller, P. F. Wilson, N. P. Bryner, M. H. Francis, J. R. Guerrieri, D. W. Stroup, L. Klein-Berndt
National Institute of Standards and Technology (NIST)
Contact: L. E. Miller, 301-975-8018, 301-590-0932 (fax), lmiller@antd.nist.gov

An indoor localization and communication project is described that proposes to use RFID tags, placed in the building beforehand, as navigation waypoints for an inertial navigation system carried by a first responder. The findings from the first year of the planned three-year project are summarized.

1. Introduction

RFID (radio-frequency identification) devices commonly are attached to persons or to moveable objects so that the objects can be tracked using fixed readers (special-purpose radios) at different locations. In this project, supported by the NIST Advanced Technology Program (ATP), we are exploring a novel application of the "flip side" of this practice based on the concept that detection of an RFID device in a known, fixed location by a moving reader provides a precise indication of location for tracking the person or moving object that is carrying the reader. The research aims to evaluate the exploitation of this concept to implement a low-cost, reliable means for tracking firefighters and other first responders inside buildings, where navigation using GPS is not reliable—indeed, the GPS signal may have been disabled temporarily to prevent exploitation by terrorists [1]. The research will also consider the use of building-related information stored in RFID devices placed at fixed on-site locations to aid the responders in their mission, as well as to describe the room layout or other context of the device, thereby minimizing the need for accessing remote databases through the communication system.

Previous research and development for indoor localization includes that of a wireless network that integrates communications, precise tracking, and data telemetry, based on ultrawideband (UWB) technology, for use in hospital and manufacturing environments [2]. In contrast, the system envisioned by this new project is intended for an environment that is potentially much less "friendly" to RF propagation—the in-building environment of first responders that may contain smoke, dust, or flames—and is intended to leverage advances in ubiquitous RFID tag technology, in combination with recent advances in miniaturized inertial sensors, to develop a low-cost tracking system that does not depend upon the stability of the RF environment over relatively large distances to derive range from precision timing. The "philosophy" of the proposed RFID-assisted system also involves reducing the dependence on RF links to external data sources by exploiting the capability of RFID tags to store critical building information for retrieval when it is needed, where it is needed.

This project is a joint effort by components of three NIST laboratories: the Wireless Communication Technologies Group of the Information Technology Laboratory (ITL) and the Fire Fighting Technology Group of the Building and Fire Research Laboratory (BFRL) in Gaithersburg, Maryland; and the Radio-Frequency Fields Group of the Electronics and Electrical Engineering Laboratory in Boulder, Colorado.

1.1 Concepts Motivating the Study

Indoor Navigation Cannot Depend on GPS. It cannot be assumed that GPS position solutions will be available to first responders in an indoor mission-critical situation. Even if the GPS signals are not blocked or obscured for tactical advantage, the reception of GPS signals inside most buildings is not reliable.

Inertial Sensors Can Track Location, Motion. In addition to, or in place of, GPS, the position of a first responder inside a building can be tracked using inertial sensors such as accelerometers and gyroscopes. Non-inertial sensors such as magnetometers and barometers can also be used in conjunction with dead reckoning to develop positions of a first responder in motion.

RFID Fixes Can Enhance the Accuracy of Inertial Tracking Systems. Inertial tracking systems inherently drift over time and produce errors in position, especially for inexpensive and lightweight systems. Corrections to the position solution at points along the path of the first responder can limit the maximum error to an acceptable level. Corrections, in the form of the insertion of known locations that have been reached, can be developed automatically by the detection of an RFID device, either by correlating the identity of the device with a table of locations or by reading the device's location from data stored on it.

1.2 Approach to the Study

At the outset of the study, the overall approach was described as follows: In addition to the RF propagation environment of buildings in emergency situations, the

research will consider several operational scenarios consisting of (1) the strategy for RFID deployment, (2) the tracking method, and (3) the options for presenting location information to the user and communicating this information to a monitoring station. The RFID deployment and tracking aspects of the scenarios to be studied will include:

The tradeoffs involved in the choice of RFID devices for this application, including cost, ease of programming, suitability for emergency environments, and data capacity.

Use of relatively few RFID location reference points to correct or calibrate an inertial navigation or other localization system to maintain sufficient accuracy during a first responder incident.

Use of multiple RFID location reference points to furnish data for tracking without the use of inertial sensors.

The emphasis will be to make maximum use of information and to leverage software to simplify hardware implementations. The presentation and communication aspects of the scenarios to be studied will include:

Informing the user (only) of position (stand-alone mode), assuming any communication is provided by a separate system.

Informing the user, other team members, and an incident commander of their positions via an ad hoc network of radio terminals that combine RFID reading and radio communication.

Providing the user with directions for safe exiting of the building.

1.3 Project Milestones and Plans

The project goals are rather ambitious. But the milestones do serve to show how the work is intended to sequence and how the three groups collaborate.

1.3.1 FY 2005 Milestones

A. Define critical parameters of firefighter localization and in-building informational requirements in typical scenarios that relate to the building RF propagation environment and to the number and placement of RFID tags in buildings, as well as the type of data to be stored on the tags. (BFRL, EEEL)
B. Evaluate inertial and dead-reckoning navigation techniques and device options (including MEMS-based sensors) regarding their accuracy, availability, and suitability for integration with an RFID reader on a small platform for location updating. (ITL)
C. Analyze the requirements for the number of RFID tags and their placement to achieve desired localization accuracies, as a function of navigation techniques and device options. (BFRL, ITL)
D. Evaluate options for RFID technologies—including both tags and readers—to use for location updating of a navigation system implemented on a small, battery-powered device similar to a handheld computer or PDA (personal digital assistant). (EEEL, ITL)
E. Establish a project web page. (ITL)[1]
F. Document/publish interim results. (ITL, EEEL, BFRL)

1.3.2 FY 2006 Milestones

A. Select RFID tag and reader technologies and develop a prototype reader for use in this application. (EEEL)
B. Develop embedded software for acquisition of data from the RFID reader and use of that data to perform location updates and to display the location on a handheld computer as well as building information derived from RFID tag data. (ITL, EEEL)
C. Conduct preliminary experiments in NIST's Large Fire Facility to evaluate the performance of RFID-assisted localization devices in structures of simple geometry. (BFRL)
D. Evaluate options for interfacing the localization device with an ad hoc wireless communication network (ITL, BFRL)
E. Document/publish interim results. (ITL, EEEL, BFRL)

1.3.3 FY 2007 Milestones

A. Develop embedded software for interfacing an RFID-assisted localization device with an ad hoc wireless communication network. (ITL, BFRL)
B. Integrate RFID reader, navigation hardware and software, and ad hoc communication system for prototype testing. (EEEL, ITL)
C. Identify test sites/buildings and conduct tests to demonstrate the operation of the prototype localization and communication system in burning or smoke-filled building environments. (BFRL, ITL)
D. Develop embedded software for directing the user to the nearest RFID-tagged exit. (ITL)
E. Document/publish final results. (ITL, EEEL, BFRL)

[1] The project web page is found at http://www.antd.nist.gov/wctg/RFID/RFIDassist.htm

2. Summary of FY 2005 Accomplishments

2.1 Firefighter Localization Parameters

Firefighter location and in-building information requirements have been developed, as representing perhaps the most demanding environment for an RFID-assisted localization and communication system. The implications for placement of tags have been identified as the survival of the functioning of the tag in the environment.[2] The number of RFID tags is still an open question, depending on a tradeoff between practicality and the desired localization resolution—in many first responder scenarios, it would be a much-needed advancement to be able to identify which room in a building a given first responder is located. This question will continue to be considered in later phases of the project.

Firefighter location and in-building information requirements may be grouped in terms of (a) building type, (b) temperature environment, (c) radio attenuation factors, and (d) desired location resolution.

Building type refers both to the building's construction, which relates to its resistance to fire, and to factors affecting communications in the building. The building type, along with classification of the building use (e.g., residential or industrial), is the primary parameter in the description of the fire event scenario. Building and fire codes classify buildings according to the type of construction and the fire resistance of the various load-bearing and non load-bearing elements, such as exterior and interior walls, columns, beams and girders, and floor construction. There are five types of building construction identified in the various codes [3, 4]:

Type I Buildings are classified as Fire-Resistive. Most of these buildings are used as high-rise office buildings, shopping centers, or residential units. They will be either reinforced concrete or structural steel. Structural members will be approved noncombustible or limited-combustible materials with specified fire resistance ratings. Any steel construction members must be protected to withstand prescribed test temperatures for fire resistance.

Type II Buildings are classified as Noncombustible. These buildings can be used for example as office buildings, warehouses, or automobile repair shops. There are three basic types of non-combustible buildings: metal frame covered by metal exterior walls (Butler Buildings), metal frame enclosed by masonry as non-bearing exterior walls, and masonry bearing walls supporting a metal roof. These metal roofs can be either solid steel girders and beams, or lightweight open-web bar joists or a combination of both with corrugated metal sheathing. The structural members are noncombustible or limited-combustible materials.

Type III Buildings are classified as Ordinary. Type III buildings are often called ordinary buildings or combustible/noncombustible. The majority of buildings probably fall into the Type III category. These buildings can be office buildings, retail stores, mixed occupancy, dwellings, or apartment buildings. These buildings usually have non-combustible bearing walls and combustible roofs. Usually, the exterior walls are concrete, concrete block, or brick. Interior, non-load bearing walls can be made of wood.

Type IV Buildings are classified as Heavy Timber. The exterior and interior walls and structural members that are portions of walls must be of approved non-combustible or limited combustible materials. Interior structural members, including walls, columns, floors, and roofs, are large dimension solid or laminated wood timbers. The exterior walls are typically masonry. These buildings exist primarily in the New England area.

Type V Buildings are classified as Wood Frame. Wood frame buildings generally are constructed in one of five methods: log, post and beam, balloon, platform, and plank and beam. A wood frame building can be used for many different purposes, such as single-family dwellings, multiple-family dwellings, restaurants, or retail stores. The major structural members are typically composed of wood and the exterior walls are combustible.

The communications factors associated with the building type include whether the building has a pre-wired communication system and whether the construction of the building is such that radio signals may not penetrate the building adequately (*e.g.*, steel or metal).

The temperature environment of a building during a first responder event is described in terms of zones that correspond to degrees of exposure to heat flux and therefore to risk of injury [5]. Tables 1–4 give examples of fires in the different temperature and heat flux zones (Zone I to Zone IV).

Radio attenuation factors are those affecting the transmission of radio signals into and out of a burning building. These factors include

Presence of water in the air, due either to combustion products or the fire suppression water. 100% relative humidity can be expected.

Smoke particulates in the air, usually carbon. Typically one gram per cubic meter.

Charged particles in the air that are ions from combustion processes.

[2] High temperature testing of the functionality of typical RFID tag systems will be conducted in FY 2006, using FY 2005 DHS funds that have been made available in part on an interest in this ATP project.

Table 1. Examples of Zone I environments.

Group, Year	Designation	Description	Ranges
USFA, FEMA; 1992 [6]	None		
Abeles Project Fires, 1980 [7]	Class 1	Overhaul, up to 30 minutes	Temperature to 40 °C Flux to 0.5 kW/m^2
IAFF; Based on Abeles Project Fires 1980	Class1	Overhaul, up to 30 minutes	Temperature to 40 °C Flux to 0.5 kW/m^2
FRDG, FEU 1995 [8]	Routine	Elevated temperature, no direct thermal radiation, 25 minutes	Temperature to 100 °C Flux to 1.0 kW/m^2
Coletta, 1976 [5]	None		
Abbott, 1976 [9]	None		

Table 2. Examples of Zone II environments.

Group, Year	Designation	Description	Ranges
USFA, FEMA 1992 [6]	Routine	One or two objects burning	Temperature 20 C to 60 °C Flux 1.0 to 2.1 kW/m^2
Abeles Project Fires, 1980 [7]	Class 2	Small fire in a room, 15 minutes	Temperature 40 °C to 95 °C Flux 0.5 to 1.0 kW/m^2
IAFF; Based on Abeles Project Fires 1980	Class 2	Small fire in a room, 15 minutes	Temperature 40 °C to 93 °C Flux 0.5 to 1.0 kW/m^2
FRDG, FEU 1995 [8]	Hazardous	Elevated temperature and direct thermal radiation, 10 minutes	Temp. 100 °C to160 °C Flux 1.0 to 4.0 kW/m^2
Coletta 1976 [5]	Routine	Fighting fires from a distance	Temperature to 60 °C Flux 0.4 to 1.25 kW/m^2
Abbott 1976 [9]	Routine	Fighting fires from a distance	Temperature to 70 °C Flux 0.5 to 1.7 kW/m^2

Table 3. Examples of Zone III environments.

Group, Year	Designation	Description	Ranges
USFA, FEMA 1992 [6]	Ordinary	Serious fire, next to a room in flashover; 10 to 20 minutes maximum	Temp. 60 °C to 300 °C Flux 2.1 to 25 kW/m^2
Abeles Project Fires, 1980 [7]	Class 3	Totally involved fire, 5 minutes	Temp. 95 °C to 250 °C Flux 1.0 to 1.75 kW/m^2
IAFF; Based on Abeles Project Fires, 1980	Class 3	Totally involved fire, 5 minutes	Temp. 93 °C to 260 °C Flux 1.0 to 1.75 kW/m^2
FRDG, FEU 1995 [8]	Extreme	Rescue, retreat from flashover or backdraft	Temp. 160 °C to 235 °C Flux 4.0 to 10.0 kW/m^2
Coletta 1976 [5]	Hazardous	Outside burning room or small building	Temp. 60 °C to 300 °C Flux 1.25 to 8.3 kW/m^2
Abbott 1976 [9]	Ordinary	Outside burning room or small building	Temp. 70 °C to 300 °C Flux 1.7 to 12.5 kW/m^2

Table 2.4 Examples of Zone IV environments.

Group, Year	Designation	Description	Ranges
USFA, FEMA 1992 [6]	Emergency	Severe and unusual, 15 to 30 seconds for escape	Temp. 300 °C to 1000 °C Flux 25 to 125 kW/m^2
Abeles Project Fires, 1980 [7]	Class 4	Flashover or backdraft, up to 10 seconds	Temp. 250 °C to 815 °C Flux 1.75 to 42 kW/m^2
IAFF; Based on Abeles Project Fires, 1980	Class 4	Flashover or backdraft, up to 10 seconds	Temp. 260 °C to 815 °C Flux 1.75 to 42 kW/m^2
FRDG, FEU 1995 [8]	Critical	Could be encountered briefly	Temp. 235 °C to 1000 °C Flux 10 to 100 kW/m^2
Coletta 1976 [5]	Emergency	Not normally encountered, may be during flashover	Temp. 300 °C to 1000 °C Flux 8.3 to 105 kW/m^2
Abbott 1976 [9]	Emergency	Not normally encountered, may be during flashover	Temp. 300 °C to 1100 °C Flux 12.5 to 208 kW/m^2

Thermal layers that could reflect or refract radio waves.

Construction materials and their various attenuation properties. Adverse effects can be from the metal facing on insulation, metal rebar in concrete, metal siding, and solar window treatments.

The desired resolution of indoor location information during a first responder event varies according to the scenario, that is, whether the building is residential or industrial. Tables 5 and 6 correlate the location resolution in meters to the accuracy in locating personnel and escape openings.

Table 5. Location parameters for residential scenarios.

Resolution in meters	Location		Escape	
	X-Y Direction	Z Direction	X-Y Direction	Z Direction
100	City Block +/-	10 floors +/-		
10	Front or rear of house	3 floors +/-	Structure +/- (Townhouse)	Floor +/-
1	Room	Floor +/-	Correct Wall	Window or Door
0.1	Location in Room	Correct Floor	Location on wall	Height of window or door

Table 6. Location parameters for industrial scenarios.

Resolution in meters	Location		Escape	
	X-Y Direction	Z Direction	X-Y Direction	Z Direction
100	Building +/-	10 floors +/-		
10	Section of Bldg	3 floors +/-	Section of Bldg	Floor +/-
1	Room	Floor +/-	Correct Wall	Window or Door
0.1	Location in Room	Correct Floor	Location on wall	Height of window or door

2.2. Navigation Techniques and Devices

The surveys and evaluations of navigation techniques and inertial navigation sensor technologies have been completed, with the major findings summarized below. On the basis of the surveys, a particular dead-reckoning module (DRM) has been selected for testing. A report containing survey details and tutorial materials on navigation has been drafted; its completion awaits testing of the selected DRM so that the test results may be included.

The most widely used navigation system today is the Global Positioning System (GPS), which enables position determination through the measurement of time delays of signals from multiple satellites in known (moving) positions; the time delay measurements are based on cross-correlating received satellite signals with local replicas to identify the signals' digital code position in time relative to the common reference. The difficulty in using GPS indoors and in urban "canyons" is that the line of sight to the GPS satellites is obscured or severely attenuated. Without four good satellite signals, the GPS position solution is inaccurate. Also, with weak signals, the GPS receiver continually loses lock and must spend an inordinate amount of time in attempting to acquire the signals.

Prior to the establishment of GPS, of course, many techniques and devices for navigation have been used. Today's navigation devices implement some very old navigation techniques, such as dead reckoning and waypoint navigation. Dead Reckoning (DR) is the process of estimating position by advancing a known position using course, speed, time and distance to be traveled—in other words, figuring out where one will be at a certain time if he holds the planned speed, time and course [10]. The usefulness of the technique depends upon how accurately speed and course can be maintained on a given "tack;" in the air and on the sea, the selection of fixed speed and course for relatively long periods of time are feasible, while on land or inside buildings the duration of the tack may need to be relatively short due to maneuvers that are required by the terrain or building layout. As illustrated in Figure 1, the uncertainty of the DR position grows with time, so that it is necessary to check the position regularly with a "fix" of some kind (perhaps an RFID tag as envisioned by this project).

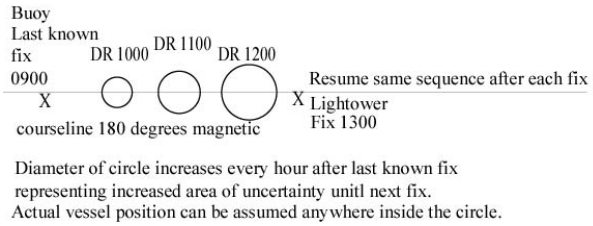

Figure 1 Dead reckoning in open ocean (from [11]).

Various systems are being proposed for "pedestrian navigation" utilizing DR techniques and small compasses. For example, [12] describes a small DR unit that utilizes a two-axis compass, a three-axis compass, or a rate gyroscope to track a walking person's heading. The heading produced by the 3-axis sensor is least affected by deviations in the person's orientation—very important in the case of first responders, who often do not walk "normally" in the course of their work. For computation of speed (actually, displacement as a function of time), the system in [12] uses an accelerometer to detect the person's steps; on average, the distance covered by a step was found to be quite consistent, even though the detection process can be rather subtle.

The outdoor performance of the system in [12] with a 2-axis magnetic compass and step detection using an accelerometer in an urban "tourist area" test is shown in Figure 2 in comparison to GPS. The scales in the figure are in meters. Although the DR positions were generally in agreement with those developed by GPS, they were often significantly different, due in part to unknown magnetic effects along the path, probably from the presence of an underground electric utility substation along the path. The standard deviation of the position area for the test was about 20 m. The authors conclude, "This case illustrates the susceptibility of magnetic compasses to localized magnetic fields, particularly in an urban environment. In these circumstances a gyroscope solution is clearly advisable."

Another personal navigation product based on dead reckoning and step counting is a DRM that integrates a GPS receiver with a magnetic compass and other sensors [13–16] and is described as follows [14]:

> The Dead Reckoning Module (DRM®) is a miniature, self-contained, electronic navigation unit that provides the user's position relative to an initialization point. The DRM® is the first commercially available practical implementation of a drift-free dead reckoning navigation system for use by personnel on foot. It is specifically designed to supplement GPS receivers during signal outages. You still know where personnel are located even when GPS is blocked by nearby buildings, heavy foliage, or even inside many structures. The DRM contains a tilt-compensated magnetic compass, electronic pedometer and barometric altimeter to provide a continuous deduced position. A microprocessor performs dead reckoning calculations and includes a Kalman filter to combine the dead reckoning data with GPS data when it is available. The filter and other proprietary algorithms use GPS data to calibrate dead reckoning sensors for a typical dead reckoning accuracy of 2% to 5% of distance traveled, entirely without GPS. Options for the system integrator include a selection of voltage input ranges, CMOS or RS232 interface, data logging, and special software functions. In addition to horizontal position data, compass azimuth, tilt (pitch and roll), and barometric altitude are available.

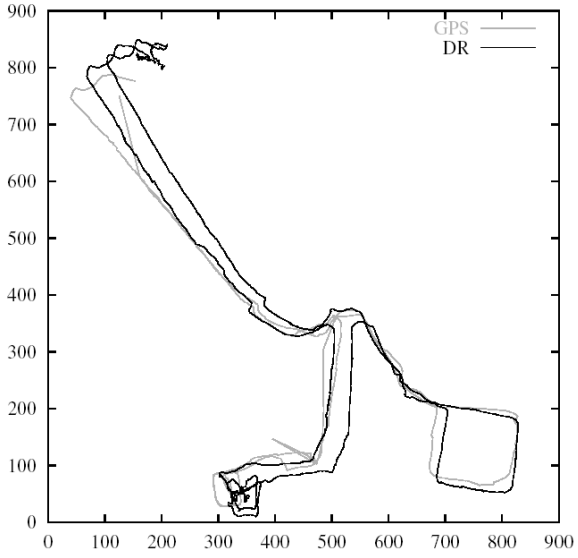

Figure 2 Positions developed by a DR system vs. GPS over a 4 km urban trail (from [12]).

For improved stability and accuracy, a version of the DRM can be obtained that includes a gyroscope. Figure 3 from [13] shows an example of the improvement in DRM performance using a gyroscope.

The relatively small size and cost of these personal navigation devices is made possible by the development in recent years of very small inertial and non-inertial sensors. For example, Micro-machined (MEMS) rate gyros based on vibration are available; a diagram of a semiconductor rate gyro based on a MEMS tuning fork is shown in Figure 4. The principle of a vibrating gyro is very simple: a vibrating object (such as a tuning fork) tends to keep vibrating in the same plane as its support is rotated. It is therefore much simpler and cheaper than is a conventional rotating gyroscope of similar accuracy. In the engineering literature, this type of device is also known as a *Coriolis vibratory gyro* because as the plane of oscillation is rotated, the response detected by the transducer (usually a piezo

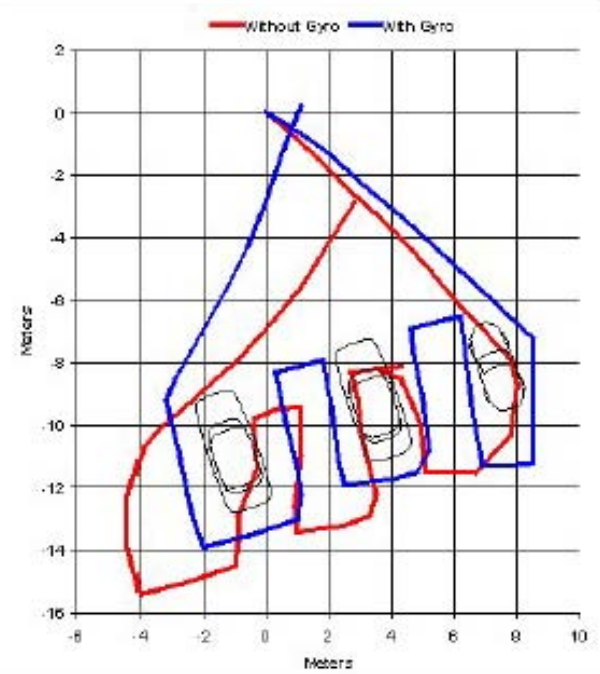

Figure 3. Manufacturer's demo of gyro-stabilized dead reckoning module improvement (from [14]).

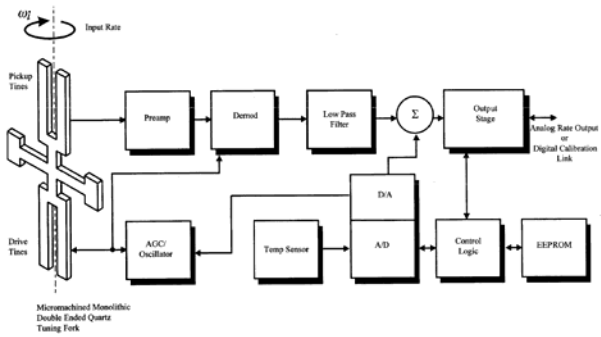

Figure 4. Block diagram of a semiconductor gyro based on MEMS technology (from [17]).

electric device) results from the coriolis term in its equations of motion. [16]

There is some question as to whether the step-counting algorithms in current personal navigation devices are sufficiently sophisticated to adapt to the irregular stepping patterns of firefighters while doing their work. Even if they are not very accurate, it is possible that they are good enough to preserve a useful track if they are periodically updated by accurate position information (see below). For the purposes of this study, ITL has placed an order for the DRM described above so that it can be tested under various conditions relevant to first responder scenarios, and eventually integrated with an RFID reader for obtaining position fixes from RFID tags.

2.3. Number and Placement of Tags

This milestone is predicated on the results of milestones A and B in that it assumes respectively that a required localization accuracy or range of accuracies has been stated and that the accuracies of various navigation techniques have been formulated in terms of the spatial density of waypoints and/or frequency of navigation fixes.

Although general information on the accuracy of potential navigation techniques for use indoors was developed, it was realized during the project that it was premature to analyze the final accuracy of an RFID-assisted inertial system at this time.

One factor involved is that the accuracy of the dead reckoning module to be tested is unknown under the conditions to be expected in a first responder scenario, particularly the effect of irregular walking on the step-counting algorithm.

Another factor involved is that there is no device available to first responders at this time that provides indoor location with any reliable degree of accuracy—the firefighting community would consider the ability just to identify the floor of the building on which a firefighter is located would be a step forward. Therefore, it is not appropriate to focus on analysis of the potential accuracy of a sophisticated system yet.

However, the project team did agree that it would be useful to evaluate the requirements for placement of RFID tags in order to indicate location from the tags alone. Using just RFID tags to derive the indoor location of a first responder has certain advantages that are known in advance. The "you are here" event of detecting a particular RFID tag in a building provides a positive indication of not only which floor of the building the first responder is on, but also which room or work area he or she is nearest to, assuming the ability to correlate the data on the RFID tag with building information or the existence of this information on the tag itself. That is, this result can be expected if there is no ambiguity in the case of the detection of more than one RFID tag.

In the industrial and supply chain applications of RFID technology, the expectation is that many tags will be within the range of the RFID reader, so that the standards for RFID devices provide protocols and procedures for resolving the signals from multiple tags. However, no attempt is made to determine the locations of the multiple tags, other than that they are all within the range of the reader.

Since the "read range" of RFID tags is very dependent on the specific RFID technology that is being used, some RFID tags and readers were procured and testing was performed to determine the read range and related parameters.

2.4 Options for RFID Technologies

The RFID technologies referred to in this milestone include not only the currently available devices, standards, and frequencies, but also any technique by which a person in a building can automatically detect that he is in a known location.

Discovery of a technique for deducing the orientation of a person at the time of detecting a device in a known location would be significant information for the study. For system studies, the evaluation of RFID technologies should include availability, cost, portability, and power consumption in addition to the functional parameters of read range and data capability.

NIST-Boulder has developed an RFID test bed at 13.56 MHz that complies with the ISO 14443 and 10373-6 standards. We have evaluated the 13.56 MHz passive tags. The tags and readers communicate via magnetic coupling which limits the read range to less than 10 cm for typical readers (simple loop antenna). More complex loop systems can be used to extend this range, as shown in Figure 5, but the practicality of such systems for first responder navigation remains under investigation. The key challenge is that because of the large wavelength of approximately 22 m, these systems operate in the near field with the magnetic field strength dropping as one over the distance cubed.

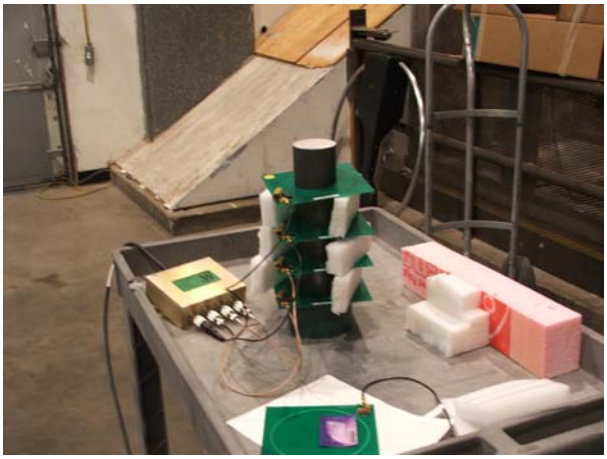

Figure 5. Loop antennas used to activate and read 13.56 MHz tags.

For higher frequencies, such as 400 MHz and 900 MHz, wavelengths are about 0.3 – 0.8 m. These systems operate in the far field. Coupling is via the electric field and the field drops more slowly (as one over the distance). This results in an anticipated read range of several meters for passive tags. It has been noted [19] that 600 MHz – 2 GHz is the best frequency band for propagation in buildings.

We have also started to evaluate an active tag RFID system that operates at 433 MHz, shown in Figure 6. This system requires an additional computer and power supply. The tags are battery powered. The tag and reader communicate using an electric field and the read range is over 30 meters. The reader can be set to recognize the tag responding with most power. This should be the tag closest to the first responder unless the battery power is low.

We have ordered 900 MHz RFID systems with software development kits. Anticipated delivery is mid-November 2005. These are passive tag systems. One uses a hand held reader that weighs only 1 kg and has a read range of approximately 3 meters (see Figure 7).

There is on-going work in the industry to optimize the performance and reduce the costs of RFID antennas and readers [20–24].

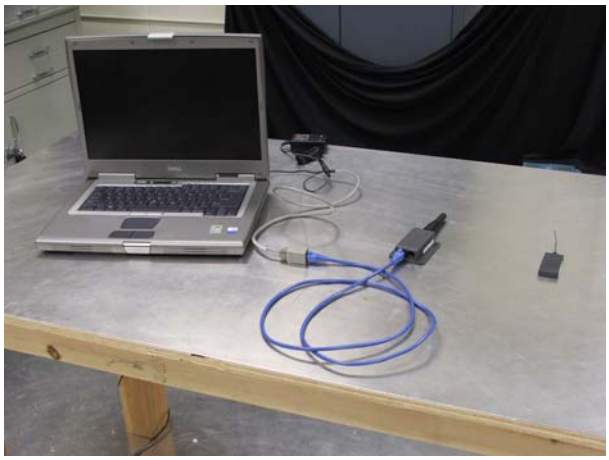

Figure 6. RFID system using active tags at 433 MHz.

Figure 7. Handheld 900 MHz reader.

3. References

[1] "Bush prepares for possible GPS shutdown," http://abcnews.go.com/Politics/wireStory?id=335054&CMP=OTC-RSSFeeds0312

[2] "Time Domain and GE Corporate R&D Announce Technology Partnership to Develop Ultra Wideband -UWB- Wireless Indoor Network," http://www.bbwexchange.com/news/archive/archived_broadband_wireless_files_2001/timedomain042401.htm

[3] NFPA A 220 Standard on Types of Building Construction, National Fire Protection Assoc., Quincy, MA 02269-9101, 2006.

[4] NFPA 5000 Building Construction and Safety Code, National Fire Protection Assoc., Quincy, MA 02269-9101, 2006.

[5] C. Coletta, I. J. Arons, L. E. Ashley, "Development of Criteria for Firefighters Gloves," NIOSH-77-134-A; GIDEP-E072-2273; Natl. Inst. for Occupational Safety and Health, Morgantown, WV, February 1976.

[6] ——, "Minimum Standards on Structural Fire Fighting Protective Clothing and Equipment: A Guide for Fire Service Education and Procurement," U.S. Fire Admin., Emmitsburg, MD, 1992.

[7] F. J. Abeles, "Project Fires, Volume 2: Protective Ensemble Performance Standards, Phase 1B," NASA CR-161530, National Aeronautics and Space Admin., Washington, DC, May 1980.

[8] J. A. Foster and G. V. Roberts, "Measurements of the Firefighting Environment – Summary Report," Central Fire Brigades Advisory Council Research Report number 61, 1994, Home Office Fire Research and Development Group, *Fire Engineers Journal*, United Kingdom, September 1995.

[9] N. J. Abbott, T. E. Lannefeld, and R. E. Erlandson, "Exploratory Development of Coated Fabric for Firefighters' Protective Clothing. Final Report. April 1974-March 1975," AFML-TR-75-72; Fabric Research Labs, Dedham, MA, July 1975.

[10] ——, http://www.deadreckoning.com/

[11] W. Newberry, "Dead Reckoning: A Skill All Navigators Should Master," Pilothouse Online at http://www.pilothouseonline.com/IS2V1_00/Lessons/main0001.htm

[12] C. Randell, et al., "Personal Position Measurement Using Dead Reckoning," *Proc. 2003 ISWC* at http://www1.cs.columbia.edu/~drexel/CandExam/Randell2003_ISWC_DeadReckoning.pdf

[13] C. T. Judd, "A Personal Dead Reckoning Module," *Proc. ION GPS '97*. Online at http://www1.cs.columbia.edu/~drexel/CandExam/DRM_ION97paper.pdf

[14] http://www.ssec.honeywell.com/magnetic/datasheets/gyrodrm.pdf

[15] ——, "DRM-III OEM Dead Reckoning Module for Personnel Positioning," Point Research Corp. brochure. Online at http://www.pointresearch.com/PDFs/DRM3.pdf

[16] ——, "Vibrating structure gyroscope," Wikipedia article online at http://www.answers.com/topic/vibrating-structure-gyroscope

[17] A. M. Madni and L. E. Costlow, "A third generation, highly monitored, micromachined quartz rate sensor for safety-critical vehicle stability control," *Proc. IEEE 2001 Aerospace Conference*, Vol. 5, pp. 2523–2534.

[18] ——, "Waypoint," Wikipedia online article at http://en.wikipedia.org/wiki/Waypoint

[19] A. E. Fathy, discussion during "Through-Wall Imaging and Sensing" July 7, 2005, session 99, 2005 IEEE Antennas and Propag. Soc. Internatl. Symp. and USNC/URSI Natl. Radio Science Mtg.

[20] S. M. Weigand, "Compact Microstrip Antenna with Forward-Directed Radiation Pattern for RFID Reader Card," July 6, 2005, session 64, 2005 IEEE Antennas and Propag. Soc. Internatl. Symp. and USNC/URSI Natl. Radio Science Mtg.

[21] L. Ukkonen, D. Engels, L. Sydänheimo, M. Kivikoski, "Reliability of Passive RFID of Multiple Objects Using Folded Microstrip Patch-Type Tag Antenna," July 6, 2005, session 64, 2005 IEEE Antennas and Propag. Soc. Internatl. Symp. and USNC/URSI Natl. Radio Science Mtg.

[22] D. Bechevet, T.P. Vuong, S. Tedjini, "Design and Measurements of Antennas for RFID, Made by Conductive Ink on Plastics," July 6, 2005, session 64, 2005 IEEE Antennas and Propag. Soc. Internatl. Symp. and USNC/URSI Natl. Radio Science Mtg.

[23] C. Cho, H. Choo, "Design of UHF Small Passive Tag Antennas," July 6, 2005, session 64, 2005 IEEE Antennas and Propag. Soc. Internatl. Symp. and USNC/URSI Natl. Radio Science Mtg.

[24] P,V, Nikitin, K.V.S. Rao, S. Lam, "Low Cost Silver Ink RFID Tag Antennas," July 6, 2005, session 64, 2005 IEEE Antennas and Propag. Soc. Internatl. Symp. and USNC/URSI Natl. Radio Science Mtg.

Extensible Software for Automated Testing of Public Safety P25 Land Mobile Radios

Julie Kub, Electronics Engineer and Eric Nelson, Electronics Engineer
Institute for Telecommunication Sciences (ITS), National
Telecommunications and Information Administration (NTIA)
325 Broadway, Boulder, CO 80305
Julie Kub: email: jkub@its.bldrdoc.gov, phone: 303.497.4607
Eric Nelson: email: enelson@its.bldrdoc.gov, phone: 303.497.4445

Abstract: One of the most prominent digital Land Mobile Radio (LMR) technologies used by public safety agencies, Project 25, is built upon an expansive suite of standards defining numerous open interfaces. As the P25 standard has matured and greater numbers of subscriber units and fixed station equipment have reached the market, increasing complaints of non-interoperability and substandard performance have arisen. In response, P25 users and manufacturers are forming a P25 Compliance Assessment Program. One element of this program requires electrical performance measurements on P25 portable and mobile radios. These tests will be conducted using an automated test software application developed at the Institute for Telecommunication Sciences. This paper describes background information on the compliance program, discusses the technical approaches taken in application development, and provides a detailed overview of the test application's functionality and capabilities.

1. Introduction

Whether responding to cataclysmic events (such as natural disasters and terrorist attacks) or performing day to day operations, it is critical that public safety personnel be able to communicate. Recognizing that cellular and land-line communication systems are often overloaded and unusable when they are needed most, public safety officials have continued to demand their own dedicated, reliable LMR systems. For a variety of reasons, public safety agencies have been slow to adopt digital modulations; most still use analog frequency modulation. However, the Federal Communications Commission's (FCC's) mandate of a transition to narrowband systems to address spectrum congestion has spurred the deployment of spectrally efficient digital modulations. In 1989 the Association of Public Safety Communications Officials (APCO) International established Project 25 (P25) to work with the Telecommunications Industry Association (TIA) to facilitate the development of standards-based narrowband digital LMR systems with open interfaces. With the maturation of the Project 25 suite of standards, the number of P25 radios on the market has recently increased dramatically. Unfortunately, operational field locations and a host of test laboratories have reported numerous interoperability and performance deficiencies. Following on the heels of September 11th, these continued reports of deficiencies and non-interoperability spurred the United States Congress to direct a formal compliance assessment program be established for Project 25 equipment procured by the Federal government and recipients of Federal grants. The following two statements are excerpted from Congressional reports:

Department of Homeland Security Congressional Report

The conferees direct the Office of Interoperability and Compatibility (OIC) to work with the National Institute of Standards and Technology [NIST] and the U.S. Department of Justice [DoJ] to require, when Project 25 equipment is purchased with such funds, the equipment meets the requirements of a conformity assessment program. The conferees further direct such a conformity assessment program be funded by this appropriation and be available by the end of fiscal year 2006 [1].

Department of Commerce Appropriations for FY06

The Committee also directs that, within this report, OLES [Office of Law Enforcement Standards] identify a process to ensure that equipment procured using Federal grant dollars complies with the requirements of the identified standard(s). At a minimum, the Office of Interoperability and Compatibility [OIC] within the Department of Homeland Security should consider working with NIST and DOJ to require that all grant dollars for interoperable communication be used for Project 25 compliant equipment that meet the requirements of a conformity assessment program [2].

Accordingly, NIST's Office of Law Enforcement Standards (OLES) working on behalf of the Department of Homeland Security's SAFECOM program in conjunction with the Institute for Telecommunication Sciences (ITS) has spearheaded the creation of a Project 25 Compliance Assessment Program. Program managers in consultation with members of the APCO 25 Interface Committee (which represents both P25 user and manufacturer interests) have identified three essential elements of a compliance assessment program,

namely, tests for conformance, interoperability, and performance. Efforts to implement performance tests have proceeded quite rapidly due to the availability of published test procedures. ITS engineers have been engaged for a number of months in a process to create the P25 Radio Performance Measurement (RPM) software suite of performance tests. This software represents a significant technical contribution to P25 radio testing. This paper focuses on the efforts to automate the performance testing element of the compliance assessment program. Interoperability and conformance testing are not discussed in this paper, because they require significant manual operations which do not lend themselves to automation.

Once software automation of the performance tests is complete, third-party independent laboratories will perform the actual tests. Many qualified laboratories such as those designated as FCC's Telecommunication Certification Bodies already have significant experience testing land mobile radios for compliance with FCC Parts 22 and 90. With the RPM software, laboratories will be able to rapidly expand their existing accreditation to include digital radio tests via NIST's National Voluntary Laboratory Accreditation Program (NVLAP). ITS will act as a technical bridge between TIA, NIST NVLAP, and participating laboratories.

2. Automated Software for Performance Testing

Performance tests are natural candidates for automation. First, uniform test protocols implemented through automated test procedures can ensure consistent methodologies amongst the various participating laboratories. Also, automation encapsulates the highly sophisticated testing techniques and equipment setup requirements of P25 digital radio testing and thus reduces the level of experience required of engineers and technicians at the independent laboratories. Finally, automated testing can either drastically reduce testing time through the use of efficient testing algorithms or enable exhaustive testing which would otherwise be prohibitive. For instance, bench tests conducted at ITS indicated that an exhaustive execution of the Spurious Response Rejection procedure required bit error rate measurements at approximately 100,000 interference frequencies. Manually conducting such an exhaustive test would be inconceivable but is readily accomplished via unattended test procedures. To promote the use of uniform test methods NIST/OLES directed ITS to develop automated test software for participating laboratories. The software suite will be provided through traditional software distribution channels. As a result, P25 radio manufacturers need not develop their own internal performance testing applications, thus promoting industry and competition.

3. Identifying LMR TIA-102 Digital P25 Measurements

Project 25 radio performance tests and accompanying requirements are documented in TIA-102.CAAA[3] and TIA-102.CAAB[4], respectively. The requirements consist of 18 receiver and 20 transmitter measurements. From these, a total of 10 receiver and 8 transmitter measurements were selected, based on their practical significance and utility, for inclusion in the performance testing component of the P25 Compliance Assessment Program. The proposed tests were reviewed by experienced LMR test engineers and accepted by the APCO 25 Interface Committee. Several test procedures were omitted since the characteristics they measured are readily evaluated by observation (e.g., audio quality and audio levels tests). Other procedures can be tested using existing FCC measurement procedures on the analog section of the radios. The 18 measurements selected for the performance test suite are being combined into one testing application, the P25 Radio Performance Measurements (RPM) tool. The tests that compose the P25 RPM tool are listed in Table 1.

Table 1. Digital P25 Automated Measurements

Receiver Measurements		Transmitter Measurements	
Section	**Procedure**	**Section**	**Procedure**
2.1.4	Reference Sensitivity	2.2.5	Modulation Emission Spectrum
2.1.5	Faded Reference Sensitivity	2.2.8	Adjacent Channel Power Ratio
2.1.6	Signal Delay Spread Capability	2.2.9	Intermodulation Attenuation
2.1.7	Adjacent Channel Rejection	2.2.12	Transmitter Power and Encoder Attack Time
2.1.8	Co-Channel Rejection	2.2.14	Transmitter Throughput Delay
2.1.9	Spurious Response Rejection	2.2.15	Frequency Deviation for C4FM
2.1.10	Intermodulation Rejection	2.2.16	Modulation Fidelity
2.1.11	Signal Displacement Bandwidth	2.2.18	Transient Frequency Behavior
2.1.17	Late Entry Unsquelch Delay		
2.1.18	Receiver Throughput Delay		

4. General RPM Hardware Requirements

The receiver measurements contained in the RPM require one or more radio frequency (RF) signal generators which emit either desired or interference type signals. These signals are combined and fed into the LMR under test by direct conduction into its antenna terminal. Most of the receiver measurements (except 2.1.17 and 2.1.18) require continued observation of a radio's bit error rate (BER) which is usually driven to a target value of 5% under a variety of simulated conditions. Significant BER averaging is required to achieve consistent results while the input signal is gradually adjusted until the target BER is obtained. A typical measurement takes 5-10 minutes while the longest receiver measurement must run unattended for several hours.

Only a handful of RF signal generators on the market are capable of generating the Compatible 4 Level Frequency Modulation (C4FM) required to perform tests on P25 receivers. Fewer still meet the phase noise requirements of TIA-102.CAAA. Two test procedures (2.1.5 and 2.1.6) also require a fading channel simulator. ITS acquired two different models of signal generator which satisfied the above requirements.

As an example of a receiver measurement, Figure 1 shows a block diagram of TIA-102.CAAA, Section 2.1.10, the Intermodulation Rejection measurement.

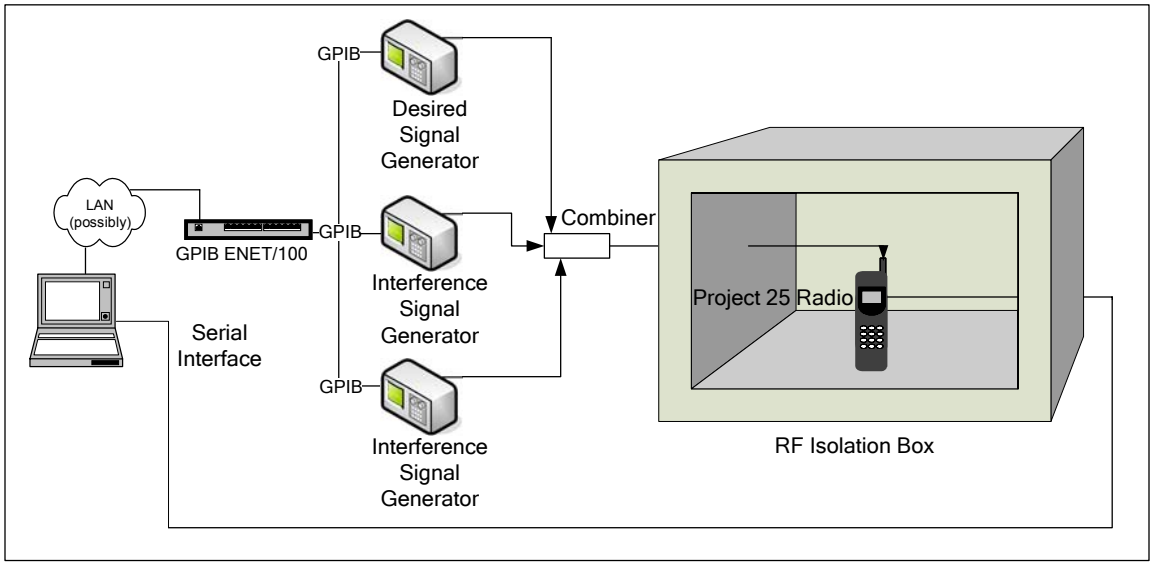

Figure 1. Example Receiver Measurement Block Diagram (2.1.10, Intermodulation Rejection)

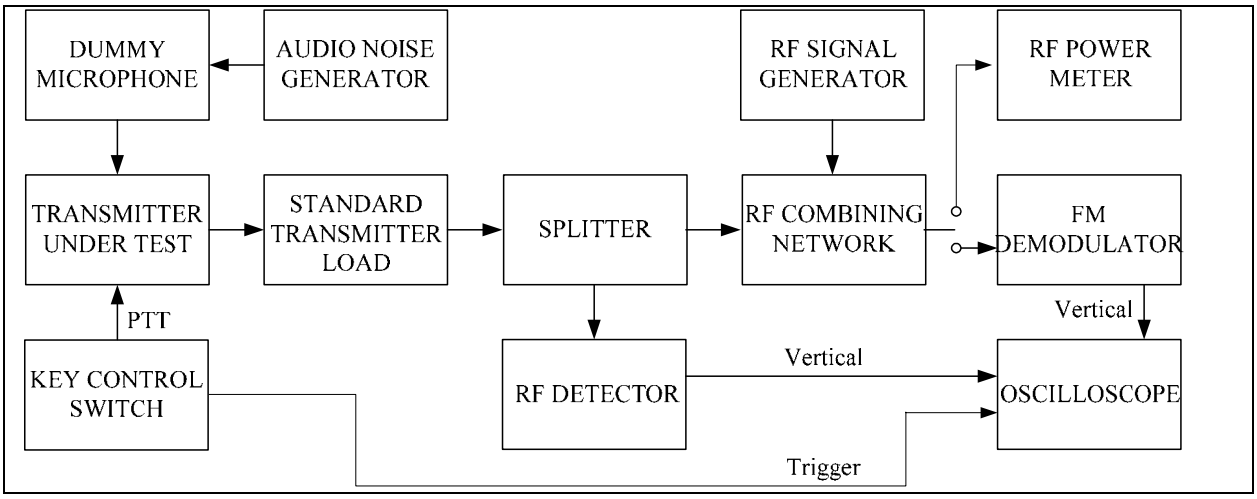

Figure 2. Transmitter Measurement Equipment without Using RTSA[3]

The automated transmitter measurements, on the other hand, are generally much less sophisticated and faster to execute (typically less than five minutes per measurement). All of these measurements can be performed using a single instrument—a real time spectrum analyzer (RTSA) with multi-domain functions. In some instances this instrument effectively replaces several other instruments as shown in Figure 2 (equipment other than the RTSA) and Figure 3 (RTSA only). Most of the transmitter tests are accomplished by configuring the device under test to transmit a particular bit sequence and asserting push to talk for a brief period of time. The RTSA captured waveform is then examined for compliance with the standards.

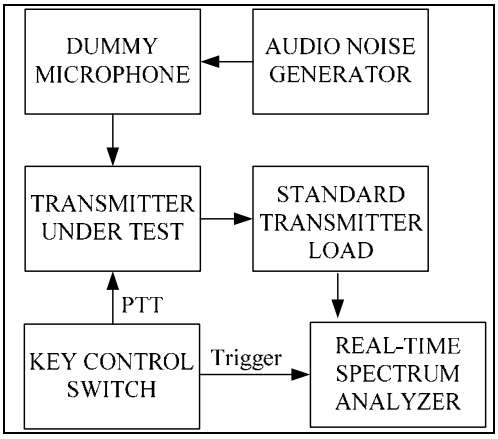

Figure 3. Transmitter Measurement Equipment Using Only RTSA

Both digital and receiver measurements are controlled via a computer running the RPM software. The test equipment, i.e., RF signal generators and spectrum analyzer, is controlled using a General Purpose Interface Bus (GPIB) to Ethernet controller. The computer's serial port is used to monitor BER data (for receiver tests) and for controlling the radios (transmitter tests). Per the recommended practice, the LMR under test will be housed in a table top RF shielded test fixture while the entire suite of test equipment will be contained in an RF shielded enclosure. TIA-102.CAAA contains detailed specifications for all eighteen transmitter and receiver measurements.

5. General RPM Software Requirements

Since the test procedures, especially the receiver tests, follow a consistent pattern, the opportunity to re-use code was apparent. Object-oriented code was used extensively throughout the RPM software for clarity, code re-use, and modularity. Every single measurement module uses a standard template containing state machines, objects, code templates and event-driven graphical user interfaces.

The RPM employs extensive data-driven design by using three databases: one for instrument commands and information, one for measurement information, and one for storing the results of all 18 measurements. The software development effort follows standard software design practices including the creation of requirements, design, and test plan documentation, and the utilization of a bug tracking database. The RPM's core measurement structure is easily extensible. Additional measurements are easily constructed from existing code with minor modifications to the measurement database and the core engine.

6. Typical Program Flow

Figure 4 shows the RPM automation state flow diagram. Of particular importance is the generic nature of the RPM software; in fact, the RPM flow chart could describe any measurement automation system that requires GPIB test equipment control and measurement automation. The flowchart shows common features of software initialization, selecting measurements to test, resetting and ensuring that all test equipment is responding, selecting the inputs for all measurements, running a batch of measurements unattended, and then performing clean-up functions. Batch measurement processing is made possible by up-front measurement procedure selection allowing several procedures to be run without further user input – a critical time-saving design feature.

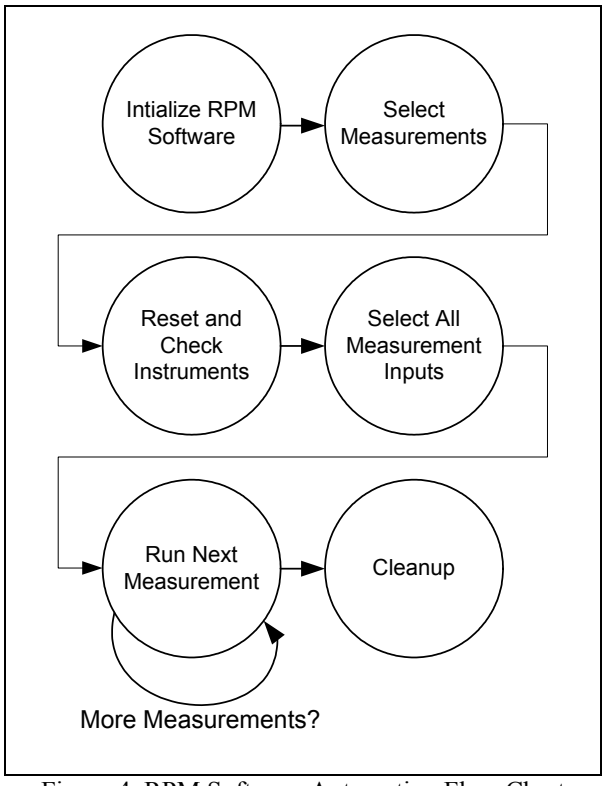

Figure 4. RPM Software Automation Flow Chart

7. The Finished Product

The RPM system contains a number of Graphical User Interfaces (GUI) which display measurement parameters or intermediate measurement data represented in graphical and/or tabular forms. A typical user interface is illustrated in Figure 5 which shows the "Intermodulation Rejection" measurement. All measurement GUI's also contain status indicators at the bottom of the display and command buttons in the lower right which enable the user to manually pause and cancel the measurement. Care was taken to ensure that "cancel" requests from any GUI are handled promptly.

8. Extensibility

The RPM system is easily extended to incorporate new test equipment and measurement modules. For instance, a digital storage oscilloscope or communications analyzer could easily be added to the program. Hypothetically, any measurement which employs GPIB controlled test equipment can be accommodated by inserting test modules into the core code.

9. Conclusion

RPM software development is nearing completion and deployment is slated to begin by the end of 2006. The software will be made available by NIST. It will be used by accredited radio test laboratories as a part of the Project 25 Compliance Assessment Program to test P25 LMRs. Use of the RPM software will minimize testing costs for manufacturers. The resultant performance test reports will be contained in a central repository which will include interoperability and conformance tests as well. Data within this repository will assist public safety procurement officials in making informed purchasing decisions. Designed using industry best practices, the RPM software is extensible and could potentially be applied to a wide array of automated testing applications well beyond its initial design scope.

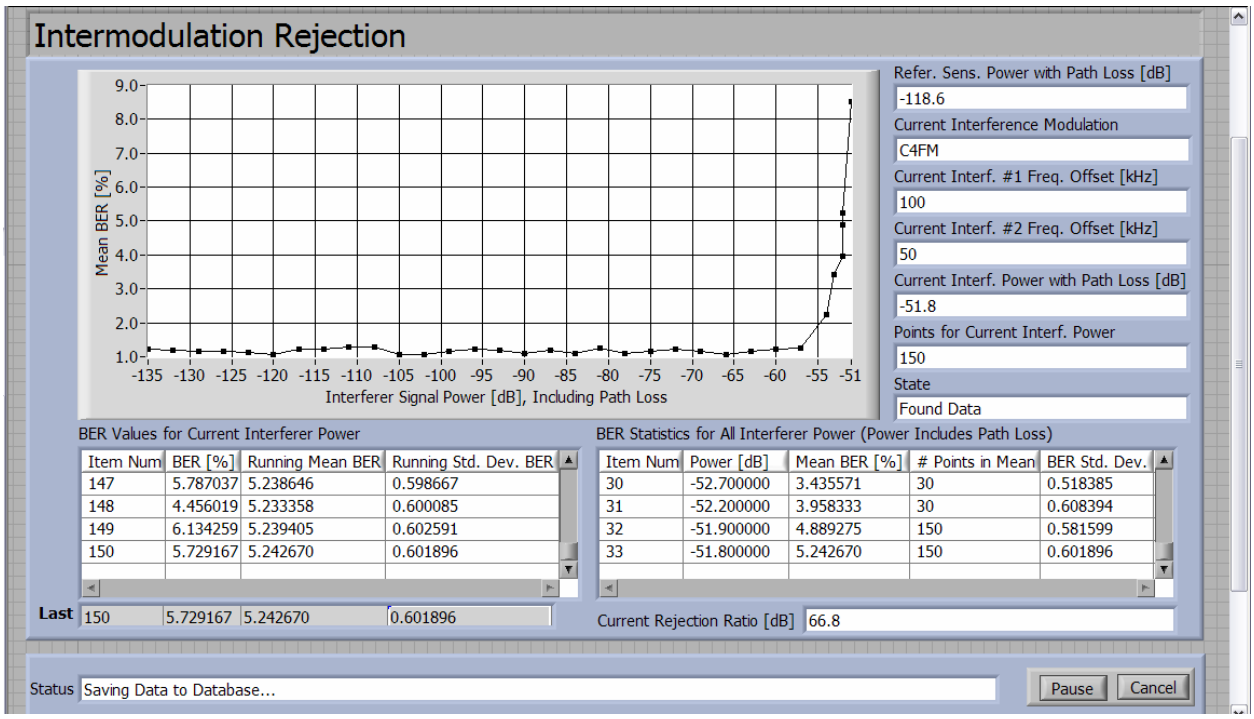

Figure 5. Digital Receiver Intermodulation Rejection Measurement GUI

10. References

[1] 109 Congress Report, House of Representatives, 23-669, 1st session, 109-241, "Making Appropriations for the Department of Homeland Security for the Fiscal Year Ending September 30, 2006, and for Other Purposes", submitted by Mr. Rogers of Kentucky, "Interoperability and Compatibility" section, September, 2005.

[2] Senate Report 109-088, "Departments of Commerce and Justice, Science, and Related Agencies Appropriations Bill, 2006".

[3] TIA-102.CAAA-A, TIA Standard, Digital C4FM/CQPSK Transceiver Measurement Methods, Sections 2.1 and 2.2, November 2002.

[4] TIA-102.CAAB-A, TIA Standard, Land Mobile Radio Transceiver Performance Recommendations – Project 25 – Digital Radio Technology C4FM/CQPSK Modulation, Sections 3.1 and 3.2, September 2002.

Flexibility in Frequency Management: The New Frequency Bill in the Netherlands

Lilian Jeanty
Radio Communications Agency the Netherlands
Phone: +31 50 5877 163
Email: lilian.jeanty@at-ez.nl

Paper unavailable at time of printing.

Why Unlicensed Use of the White Space in the TV Bands Will Not Cause Interference to DTV Viewers

Michael J. Marcus
Marcus Spectrum Solutions
+1-301-229-7714
mjmarcus@alum.mit.edu

Paul Kolodzy
Kolodzy Consulting

Andrew Lippman
MIT Media Lab

Abstract: On May 13, 2004, the Federal Communications Commission approved a Notice of Proposed Rulemaking (NPRM) proposing to allow a new generation of wireless devices to utilize vacant television channel frequencies in each market. This so-called TV band "white space" consists of frequencies that are allocated for television broadcasting but are not actually in use in a given area. This paper explains how the concept proposed by FCC, with minor modifications, can be made to work without causing harmful interference to DTV reception

The FCC's proposed rules[1] in Docket 04-186 are intended to make way for unlicensed spectrum technologies, such as Wi-Fi, to utilize the prime TV band spectrum to offer wireless broadband services. Wi-Fi technology has become very popular at higher frequencies, and has had a positive impact on the growth of broadband services. However, the bands used for Wi-Fi do not have appropriate radio propagation characteristics to serve low population densities. Lower-frequency spectrum, such as that used for TV broadcasting, is capable of traveling longer distances at a given power level, and can better penetrate obstacles such as buildings and trees.

The FCC's proposal would promote both spectrum efficiency and wireless broadband deployment. The TV band has been called a "vast wasteland" of underutilized spectrum. Even after the completion of the DTV transition – and the reallocation of TV channels 52-to-69 – an average of only seven full-power DTV stations will be operating on channels 2-to-51 in the nation's 210 local TV markets. Only a fraction of the 294 MHz of prime spectrum allocated to DTV services will actually be utilized in most markets.

Thus, the proposed use of "white space" TV channels could have a particularly great impact on the growth of information services in rural areas, where such empty channels are readily available. In urban areas, where less "white space" is available, this spectrum would also be useful because of the great demand for wireless broadband services and because of the ability of the TV band spectrum to penetrate buildings and objects within buildings better than the higher bands.

The FCC was clear in this NPRM that any devices certified to operate in the TV white spaces would be required to use *new "smart radio" technology* that would not interfere with television reception. Nevertheless, the National Association of Broadcasters (NAB) and other broadcast industry representatives, in comments filed at the FCC and in communications with Congress, have objected to the FCC's proposal, claiming that unlicensed devices operating on vacant channels in the TV band would cause harmful interference to television broadcasts and other uses of licensed TV band channels.

This Issue Brief responds to the broadcast industry's allegations, addressing each of the industry's concerns about interference. The paper concludes that interference-free unlicensed use of the white space *is* practical with today's technology. While some of the issues raised here are novel, the FCC as an "expert agency" should be able to handle them as it handles other cutting-edge spectrum problems. Indeed, the FCC is required by statute to avoid harmful interference with licensed TV broadcasts – and its NPRM describes several different ways to protect the dwindling number of over-the-air TV viewers from interference, as described below.

Unlicensed Devices: 350 Million and Booming

Unlicensed devices have been authorized by the FCC since 1938. A Consumer Electronics Association study quoted by the FCC estimates that there are over 350 million unlicensed devices in the US and that annual hardware sales are in the multibillion dollar range.[2] The earliest unlicensed devices were remote controls for radio receivers. Today's unlicensed devices range from the ubiquitous cordless telephone to garage door openers to home security systems to Wi-Fi wireless local area networks.

All of these systems comply with general rules established by the FCC to ensure that they do not cause interference to licensed systems and Federal Government systems.[3] Some unlicensed devices operate at very low power so they can coexist with higher power licensed users in the same band,[4] while others (e.g., Wi-Fi) operate in bands that are largely devoid of licensed users.[5] Before a new model of unlicensed device can be sold, it must be authorized--that is, it

must be tested by a third party and shown to comply with technical standards established in FCC Rules.[6] The FCC enforces its technical rules for unlicensed devices through both this equipment authorization program and through its statutory jurisdiction over the marketing of devices "capable of emitting radio frequency energy...in a sufficient degree to cause harmful interference to radio communications."[7]

The FCC's technical rules have been primarily focused on preventing interference with licensed users. Unlike some other governments, the FCC has not attempted to steer the market by mandating specific services or technologies. This light-handed regulation has enabled a dynamic market for unlicensed devices to develop, as innovators bring to market new devices for new applications. Perhaps the best known example of this dynamic innovation on unlicensed bands is the explosive growth of Wi-Fi technology. The Telecommunications Industry Association estimates that sales of Wi-Fi equipment in 2004 reached $4.35 billion, and predicts spending on Wi-Fi infrastructure equipment will increase to $7 billion in 2008, a 12.6 percent annual increase.[8] The development and popularization of Wi-Fi technology was built on a 1985 FCC decision[9] to allow unlicensed devices in three bands--then best known for being the "home" of microwave ovens--provided they used "spread spectrum" technology to minimize interference.

The FCC Proposal for Unlicensed Sharing of TV Spectrum Without Harmful Interference

The Commission's May 2004 NPRM proposed to allow unlicensed devices to operate on unused TV channels, often called "white spaces." As the FCC noted in its NPRM, this spectrum would be ideal for unlicensed broadband because it has better radio propagation characteristics than the present Wi-Fi bands and can tolerate higher power devices without causing interference. These characteristics allow wireless broadband providers to achieve better-quality coverage of larger areas using less infrastructure, significantly reducing the cost of broadband deployment. A recent study by Intel confirms this, showing that the capital costs of covering a rural area with wireless broadband service in the TV band would be one-fourth those needed to achieve the same coverage using licensed MMDS spectrum in the 2.5 GHz band (which sits adjacent to the current unlicensed "Wi-Fi band" at 2.4 GHz).[10]

The NPRM proposes unlicensed operation under one of three alternative schemes intended to prevent interference to television reception:

I. *"Listen-Before-Talk" (LBT)*: Sensing the presence of TV signals by the unlicensed device in order to select channels not in use. This concept, also described as dynamic frequency selection (DFS), has already been adopted by the International Telecommunications Union (ITU) and the FCC for sharing of the 5 GHz spectrum between unlicensed systems and military radar.[11] Technical protocols to avoid interference are negotiated between industry and the military.

II. *"Geolocation/Database"*: Location sensing and consultation with broadcast database. In this scheme, an unlicensed device would contain location-sensing technology, such as a Global Positioning System (GPS) receiver. The device would cross-check its own location with an internal database of TV transmitter locations in order to verify that it was a minimum distance from a TV transmitter.

III. *"Local Beacon"*: Reception of a locally transmitted signal that identifies which TV channels may be used in the local area for unlicensed use. In this scheme, low power local signals, possibly controlled by local broadcasters, would indicate directly which channels were free for use.

The FCC NPRM proposes possible use of any of these methods as acceptable ways of avoiding interference to licensed broadcast users, and recognizes that the final rules might only allow for one or two of these independent alternatives. The remainder of this Issue Brief will discuss basic technical issues that have been raised in the FCC proceeding and then specific points made by the broadcast industry lobby in recent communications with Congress and the FCC.

I. Broadcaster Interference Concerns are Unfounded or Readily Avoidable with Established Technologies

This section will address basic technical issues associated with the three alternatives. The proponents of this NPRM, including academics and equipment manufacturers, have shown in their comments that any of the three alternatives may be both effective and practical. While the original FCC proposal might not have been flawless, the remaining issues can be resolved through the normal rulemaking process at the FCC. Indeed, this is why Congress adopted the Administrative Procedures Act[12] in order to have a give and take between regulators and concerned parties

before rules are adopted reflecting the overall public interest.

A. Listen-Before-Talk (LBT) Alternative; Avoiding the "Hidden Node" Problem

The broadcast interests have focused much of their concern about the NPRM on alleged vulnerabilities in the *LBT* alternative (Alternative I above), in which unlicensed devices must first "listen" and sense the presence of TV signals in the area before transmitting. They point out that, as shown in Figure 1, an unlicensed device could be in the shadow of a building and be shielded from the TV signals, while a TV antenna at the top of the building might get a good signal.[13] This is known in the technical literature as the "hidden node" problem. Indeed, studies have shown that in both urban and rural areas, where buildings and terrain serve as obstacles to TV signal penetration, there exist many "shadow" spots in which TV signals may be weakened or totally diminished.

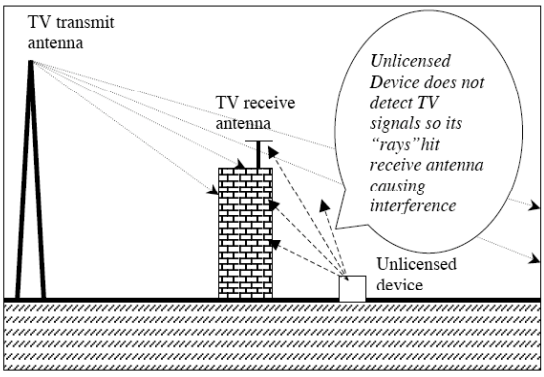

Figure 1 - The "Hidden Node" Problem

Therefore, the broadcast interests claim that unlicensed devices using this alternative are likely to miss detecting TV signals due to shadowing, and thus will cause interference to nearby TV receivers that have adequate signal strength.

The comments of the broadcast industry (and even the FCC's NPRM) assume that the detector part of the unlicensed devices in the *LBT* alternative would be about as sensitive to radio-frequency emissions as are normal TV receivers. But this need not be the case. Research presented at a February 2003 FCC-sponsored seminar demonstrated that a detector optimized for a specific class of signals (e.g., TV signals) can be orders of magnitude more sensitive than a normal receiver.[14] The Commission had previously taken note of this research in its NPRM on cognitive radio,[15] but inexplicably did not address it in this unlicensed NPRM. Similarly, the reply comments of the broadcast community have steadfastly ignored the applicability of this technology, which was mentioned repeatedly by various parties in the comment phase of the FCC rulemaking.[16]

It has also been pointed out in the comments that cooperative sensing of TV spectrum by multiple unlicensed devices could, in effect, improve sensitivity of TV signal detection significantly. Such cooperative sensing can be used in conjunction with very sensitive detectors for even more sensitivity gain.[17]

The use of very sensitive receivers could solve the hidden node problem. *The FCC could simply set a sensitivity value for detectors that would give a high confidence that usable TV signals would not be missed*, and then verify during the equipment authorization process for each model of unlicensed device whether that sensitivity level was met.

B. Geolocation/Database Alternative: Need to Keep FCC Data Up to Date

The broadcast interests also raise concerns about a second alternative means to avoid interference with TV reception on nearby channels: Geolocation and automated checking against a database of frequency assignments (Alternative II). Broadcasters have pointed out that geolocation systems such as GPS do not generally work indoors and hence could not reliably determine location. They also point out that the FCC databases on broadcast stations are not 100% accurate and are sometimes slow in catching up to transmitter frequency location changes – a more common problem now during the DTV transition.

We acknowledge the validity of these comments, but note that all of these concerns can be addressed with minor modifications to the proposed rules. The final rules should require that unlicensed devices must make iterative geolocation checks within a specified time interval in order to continue transmitting on a given frequency.

With respect to the broadcaster claims about the reliability of geolocation technologies, it is important to note that there are advanced GPS technologies used in some cellular telephone systems that actually *do* work indoors.[18] Furthermore, once the DTV transition is complete, it will become technologically feasible to conduct indoor geolocation using multiple DTV signals, instead of the satellite technology used in current GPS systems. Indeed, geolocation could even become a new product for broadcasters.

In regards to the accuracy of FCC transmitter databases,

outdated, allowing manual data entry problems to compromise the accuracy of transmitter location information. We call upon Congress and the FCC to recognize that such technology issues limit the potential of the multibillion dollar industries the FCC regulates and upgrade FCC databases so that they can be viewed as highly reliable. Regardless, if Congress mandates a "hard date" for the end of the DTV transition as it is expected to do, spectrum use will become more stable and the problems of updating the present FCC systems will become manageable.

C. Local Beacon Alternative: Control Signal Rules Can Avoid False Positives

With respect to the *Local Beacon* alternative (Alternative III above), the broadcast interests point out that the NPRM did not specifically propose what type of short-range radio signals should be used to broadcast channel availability information. Absent specific rules, a long-range transmitter might indicate availability of a certain channel and be received in an area far away, where that channel is not really available. For example, a signal transmitted in the AM broadcast band could have a range of hundreds of miles at night, and would be inappropriate for carrying information about which empty TV channels could be used in a given area. We agree with the broadcast interests on this point, but the problem could be simply resolved by rules specifying that the radio channel used to convey TV channel availability information must have a range comparable with the geographic validity of the channel availability information.[19]

D. Channel Availability

Some broadcast interests have questioned whether there will be significant channel availability for unlicensed use in major urban areas during the DTV transition. This concern is unwarranted. Even in urban areas, where there are fewer unused channels, there is likely to be substantial channel availability during the transition. Also, as Intel has argued, just because a particular channel may not be available throughout an urban region, doesn't mean it won't be available in *parts* of an urban area. Furthermore, the issue of channel availability during the DTV transition is likely to be short-lived. It now seems likely that the DTV transition will be ended by a date certain in the not too distant future – and the transition issue will simply go away.

Most importantly, there is no doubt that in rural areas—where unlicensed access to the TV band white space would make the most difference for affordable broadband deployment—there *is* spectrum available now and there will be for the foreseeable future. The proponents of this proposal do not seek a guarantee on how much spectrum will be available in a given location at a given time, and are willing to take their risks with the basic FCC proposal and their own analysis.

II. Other Concerns Expressed by Broadcast Community and Responses from NAF *et al.*

The broadcast industry has vehemently opposed the NPRM with multiple allegations that the proposals would cause serious harm to broadcast reception, cable television (CATV) reception, and to wireless microphones used in broadcast program production.[20] These allegations are addressed in turn below. The order of discussion here follows that of the April 8, 2005 letter sent by a broadcast industry consortium, the Coalition for Spectrum Integrity, to Senate Commerce Committee Chair Ted Stevens (R-AK).[21] After the discussion of these points, we address the issues raised in a recent web-based video from the broadcast lobby.

A. "Interference to 73 Million TV Sets"

The FCC has previously noted that only a steadily declining minority of households with televisions are actually dependent on over-the-air signal reception, and that more than 85% of American households with televisions subscribe to cable or satellite services,[22] and thus could not possibly be affected by interference from nearby unlicensed devices.[23] Nonetheless, the broadcast lobby asserts that permitting unlicensed broadband devices to operate on vacant TV band frequencies will cause a range of interference problems. The industry commissioned a Canadian laboratory study to corroborate these claims.[24] However, the results produced by the study were created under unrealistic conditions, such as certain combinations of channels and antennas pointing directly at each other. (This study also implicitly introduces the broadcast lobby's trick of using ultrawideband transmitters, permitted by a loophole in the original FCC proposal, to simulate the proposed unlicensed devices. This tactic forms the basis of a lobbying video released by the broadcast industry, discussed in Section I, below.)

B. DTV Disruption Issue

Broadcasters have claimed that implementation of the proposals would create consumer confusion and delay the penetration of DTV receivers needed to reach the 85% consumer take-up threshold mandated in current law before broadcasters would be required to cease analog transmissions. There is no evidence for this

assertion. Concerns have also been raised that uncertainty about this rulemaking might make small local stations delay making final channel selections and converting to DTV. However, it now appears likely that the DTV transition end-date will be mandated, rendering this issue moot. Congress is expected to pass legislation this year that will end the DTV transition by a date certain, as well as to subsidize digital-to-analog converters and an education campaign aimed principally at the 15% of households still relying on over-the-air reception.

The broadcast community's statement that unlicensed devices may cause "interference to newly purchased DTV receivers, which may cause consumers to return their new TV sets," similarly lacks a factual basis. Today's DTVs are far more capable of handling and rejecting any potential interference than older analog sets, which are sus
impairments that pass through directly to viewers in the form of ghosts, snow, and interference patterns in the video display. To suggest that new DTVs are somehow more susceptible to potential interference than other TVs is questionable logic.

C. Public Safety Interference

The *Geolocation/Database* and *Local Beacon* alternatives in the FCC proposal use local information, such as location and databases of facilities, in deciding what channel to use. Thus unlicensed systems using these techniques could readily avoid channels 14-to-20 in the handful of markets in which they are used for public safety. The *LBT* alternative requires more complexity to avoid public safety use of channels 14-to-20 since lower power, intermittent public safety communications are harder to detect than high power, full time TV broadcasting. However, technology already exists that allows unlicensed devices to detect and avoid military radar – which is a far harder task that detecting public safety communications. The FCC can solve this problem simply by requiring a long listening period on public safety channels before they can be declared vacant.[25] Similarly, the FCC can decide to require that unlicensed devices operating on certain frequencies include the ability to recognize a priority-in-use signal transmitted by public safety systems.

D. Newsgathering and Sports Programming Production

Although not generally known, broadcasters and certain other entities are allowed to use vacant TV channels for "low power auxiliary stations" (e.g., wireless microphones) with nominal licensing under the provisions of federal regulation.[26] While this use is officially licensed, this spectrum has not been auctioned and it bears many similarities to unlicensed use except that it is reserved for a narrow group of eligible devices. These devices are used at studios, but are sometimes used at sports events and other outdoor news events.

The broadcast interests raise concerns that the wireless microphones used by broadcasters on vacant TV channels might receive interference from unlicensed devices using the *LBT* alternative. While the FCC minimized this problem in the NPRM,[27] it is a difficult problem to solve in a manner that is transparent to existing users of such wireless microphones because the microphones operate at a lower power, do not necessarily have signal formats enumerated by regulations, and do not have a formal channel plan.

But it is not at all clear that such devices should continue to have exclusive access to this spectrum. The continued exclusive access of this small group of devices to large blocks of valuable spectrum for very occasional use, independent of marketplace forces, is anachronistic and inconsistent with spectrum policies enacted by Congress and implemented by the FCC in the past two decades. The FCC should perhaps revisit why broadcasters and the narrow group of eligible entities specified in FCC regulations are granted *sole access* to the "white space" spectrum in the TV band for a use that does not involve broadcasting directly to the public. When these policies were adopted decades ago, there was no other alternative to allow use of this "white space" except manual coordination among a small group of broadcast licensees. However, today's technology has increased the demand for this type of spectrum and permits cognitive radio alternatives such as those in the NPRM. Why should wireless microphones not be treated as unlicensed devices?

Even if this anachronistic use of the white spaces is continued, however, the *Local Beacon* scheme would protect wireless microphones, as local broadcasters would control the signals indicating which channels were available in a given area at a given time. There are also compromises available that could protect users of such microphones and allow the proposed unlicensed use: the FCC could, for example, adopt a transition plan that exempts unlicensed devices from certain TV channels for a transition period. Following this period, it could then grant full interference protection to eligible wireless microphone users that transmit a low power beacon signal in the vicinity of an operating wireless microphone, and having a comparable coverage area to that microphone, indicating which TV channel the microphone was using.[28] In this way, the broadcasters would have preferential (but not sole) access to the TV band.