RF Ranging for Location Awareness

Steven Michael Lanzisera
Kristofer Pister

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2009-69
http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-69.html

May 19, 2009

RF Ranging for Location Awareness

by

Steven Michael Lanzisera

B.S. (University of Michigan, Ann Arbor) 2002

A dissertation submitted in partial satisfaction of the
requirements for the degree of

Doctor of Philosophy

in

Engineering – Electrical Engineering and Computer Sciences

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

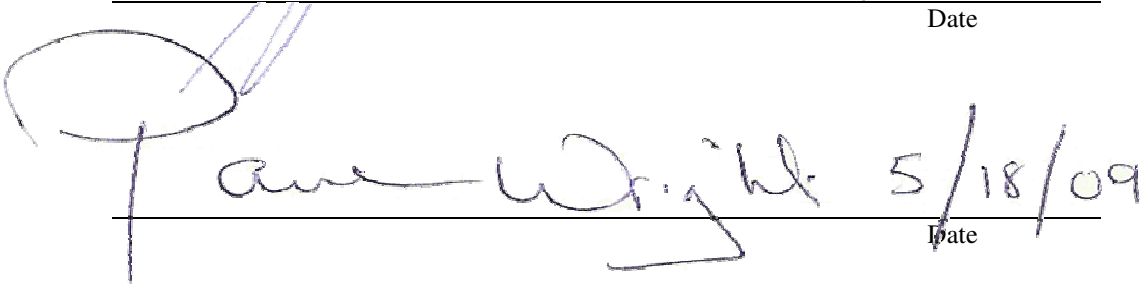
Professor Kristofer S.J. Pister
Professor Jan M. Rabaey
Professor Paul K. Wright

Spring 2009

The dissertation of Steven Michael Lanzisera is approved.

Chair Date

Date



Date

University of California, Berkeley

RF Ranging for Location Awareness

Steven Michael Lanzisera

Abstract

RF Ranging for Location Awareness

by

Steven Michael Lanzisera

Doctor of Philosophy in Engineering – Electrical Engineering and Computer Science

University of California, Berkeley

Professor Kristofer S.J. Pister, Chair

Wireless sensor networks provide an opportunity to improve performance in areas ranging from energy efficiency to industrial processes to scientific research. Many applications require awareness of sensor location, but autonomously determining device location has proven to be challenging. This localization problem can be divided into two parts: measuring relationships between nodes, and then using these relationships to estimate location. Most work on the first part has measured the RF received signal strength as a surrogate for range resulting in poor location accuracy. Several other methods have been studied with varying performance and limitations. The second part has received significant research attention resulting in several good algorithms.

This work considers the first part of the localization problem and discusses RF time of flight ranging for location awareness in local area networks. A roundtrip RF time of flight ranging method for narrow-band radios is presented that successfully deals with the many error sources that cause RF based ranging methods to suffer from poor

accuracy and high system complexity. This method has been implemented on a custom software defined radio platform and a network of these devices has demonstrated meter level location accuracy.

Professor Kristofer S.J. Pister

Dissertation Committee Chair

Contents

ACKNOWLEDGEMENTS

My time at Berkeley has been filled with countless people who have helped me along the way. There is no way to acknowledge them all, so please forgive me if I left someone out.

The faculty at Berkeley were always interested in new ideas and helping me understand old ones, and I would particularly like to thank a few for their mentorship. Kris Pister has provided tremendous freedom with the right amount of guidance and vision to push me beyond my own abilities. I would also like to thank Bernhard Boser for his advice and insight over the years. J. Rabaey and P. Wright have provided much feedback and advice regarding this dissertation. I also appreciate R. Howe, M. Maharbiz, B. Gilchrist, and K. Najafi for contributing so much to my views on research and beyond.

My fellow graduate students and researchers throughout the campus, you provided that perfect combination of diverse knowledge and commiseration that makes graduate school a great experience. Sarah, Axel, Ben, Anita, Chinwuba, Matt, Brian, Ankur, Al, and Subbu, thank you for the countless white board discussions, coffee breaks, beer breaks, and general good times in 471 Cory and elsewhere. Thanks also to my students and fellow teachers at San Quentin for such an enriching experience.

Those closest to me often get the least recognition, but they deserve the most. To my friends outside of UCB, thank you for your friendship and perspective on life on the outside. I would like to thank my parents because they instilled the value of education from an early age, and I wouldn't have made it this far without them pushing me along. Chris, your encouragement has meant more than you know. Bill, thank you for being a great friend and the source of years of good times, advice and procrastination. Most of all, I would like to thank my wife, Kristi, for her love, patience, and support. She has provided much needed balance in my life, and I am far happier and productive as a result.

Chapter 1

Introduction

1.1 MOTIVATION

Location aware wireless local area networks can determine the location of the constituent wireless nodes autonomously in addition to being capable of data communication. This combination of capabilities promises to enable applications ranging from tracking inventory in factories to locating equipment in hospitals to determining the geographic position of devices after deployment. Determining location of a device is called localization, and the localization problem is divided into two main parts. The first phase involves measuring a relationship between nodes (distance, angle, RF received signal strength), and the second phase uses these relationships to estimate location [2]. The second phase has been widely studied, and a number of good algorithms have been developed [3]. The primary area for continued research in the second phase involves determining location when some measurements are highly erroneous [4], but this second phase is not the topic of this dissertation. The first phase has seen a variety of solutions including ultrasonic time of flight (TOF) ranging, radio frequency (RF) TOF, and RF received signal strength (RSS), and these solutions have advantages and limitations important to the localization problem. Currently these methods do not provide accurate range estimation while also being compatible with the low cost radios used in wireless sensor networks and other wireless local area

networks. The work presented here considers solutions to RF ranging using narrowband radios like those typically used in wireless sensor networks.

1.2 Research Goals

The wireless estimation of range between RF devices is challenging even with the most capable radios, and ranging methods developed for the simple radios in wireless sensor networks have not provided the accuracy required for many applications. The goal of this work is to understand the performance capabilities and limitations of an RF ranging system that is compatible with local area wireless standards and to demonstrate a system capable of accurate ranging in the environments used by these networks. This work primarily considers the most limiting wireless networking standard, the IEEE 802.15.4 standard personal area networks, in order to demonstrate how much can be achieved with greatly limited resources [5].

1.3 Ranging Accuracy and Location Accuracy

Applications require location accuracy, and this is measured in terms of difference from estimated location to true location. The system under consideration here is a ranging system, and a ranging system is specified with a particular ranging accuracy. Ranging accuracy is measured in terms of the difference between the estimated distance between two nodes and the true distance, and it is important to understand the relationship between ranging accuracy and location accuracy. Localization algorithms and network geometries differ in how ranging accuracy translates to location accuracy, and many range based localization methods have been presented [3]. In order to address the link between location and range accuracy, we apply a common method of range based location estimation: the maximum likelihood estimate (MLE) of the

location based on a set of range estimates. The MLE of the location is found by calculating the probability density function (PDF) of the location based on each range estimate, multiplying the PDFs together for each range estimate, and finding the point where the resulting joint probability is maximized. Consider the case where the PDF of the location given a range estimate, r_{est}, is given by $f(r_{est}|r_{true})$. If n independent range estimates ($r_{est_1}, r_{est_2}, \ldots, r_{est_n}$) are used to find the MLE of the location, then the joint probability distribution of the location is given by the product of the individual PDFs,

$$f(\{r_{est_i}\}|l) = \prod_i f(r_{est_i}|r_{true_i})$$

where l is the location. When $f(\{r_{est_i}\}|l)$ is maximized, the corresponding location is the MLE [6-9]. The maximum likelihood estimate is the same as a minimum squared error solution if $f(r_{est}|r_{true})$ is well modeled by a zero mean normal distribution, and the minimum squared error solution is commonly used as well [10]. The left part of Figure 1.1 show the results of a random simulation of one simple 2D case when $f(r_{est}|r_{true})$ is normally distributed with parameters ($\mu=r_{est}$, σ), and hence the maximum likelihood

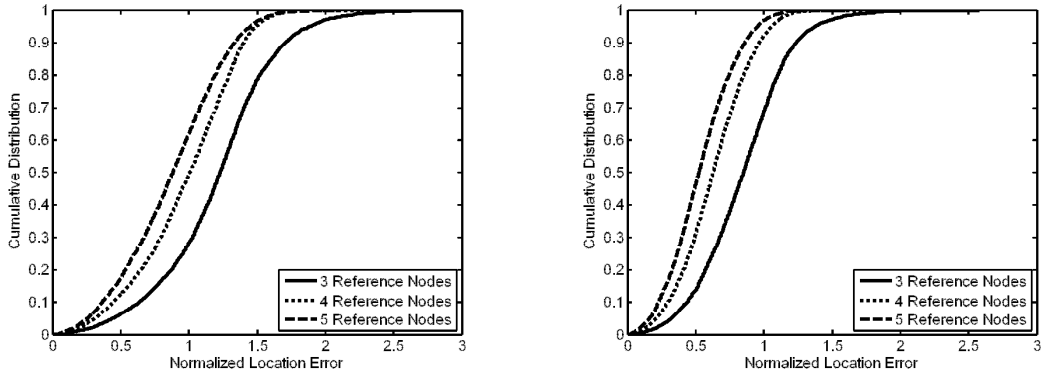


Figure 1.1 Cumulative distribution of location error normalized by the root-mean-square ranging error (left) and the maximum ranging error (right).

estimate is that same as a minimum squared error solution. In figure 1 the cumulative distribution function (CDF) of the location error normalized to the root mean square (RMS) ranging error is plotted when there are 3, 4 and 5 reference points. The value of the CDF represents the probability that the normalized location error will be less than shown on the x-axis. In the right part of Figure 1.1 the CDF of the location error normalized to the worst case ranging error is plotted. When more than 3 reference nodes are available, performance improves significantly especially when compared to the worst case ranging error. From this simulation two conclusions result: 1) increasing the density of nodes with known location is important for improving accuracy; 2) ranging accuracy and location accuracy are very similar. Although the location accuracy can be better or worse than the ranging accuracy depending on the conditions and localization algorithm used, we will assume that location error is equal to the ranging error for simplicity.

1.4 REQUIREMENTS

This section will provide a few example applications and some specifications that can be loosely derived from these applications. The applications under consideration here are asset and personnel tracking, network device localization after deployment, and building security.

1.4.1 Asset Tracking

Determining the location of people and objects in near real time is seen as the largest application of localization systems. In the hospital environment, equipment, staff and patients could all be tagged to increase efficiency and safety. It is common for hospitals to own many extra pieces of equipment in hopes of ensuring that the appropriate items can be located and used quickly. Despite this preventive measure, much time is wasted

searching for equipment. Because wasted time is so costly in terms of both dollars and care, this environment would benefit significantly by location aware devices. Not everything in a medical facility must be monitored with tight latency requirements, but short latency updates of specific items are required. Accuracy must be sufficient to ensure that the correct room is shown almost all of the time. Given that a typical hospital room is about 4 by 7 m, accuracy of better than 1.5 m ensures the correct room is indicated 50% of the time. Alarms or query targets must be localized within a few seconds. It is expected that at least one device will be in each room, but there may be several devices per room. In order to ensure enough connectivity for localization with a single device per room, a range of 15 m is required.

1.4.2 Large Network Deployment

A primary cost of deploying a large scale wireless sensor network is the installation of nodes and recording the locations of these nodes. Localization systems can reduce this cost by determining the locations of devices after deployment. Latency requirements are minimal in that it is acceptable for the initial network configuration to take hours to complete. The scale of many industrial campuses requires long ranges, possibly in excess of 100 m, but accuracy requirements depend on the location of the device. For example, devices outdoors can be localized with less accuracy and longer range, and indoor devices are more densely populated and may require 1.5 m accuracy.

1.4.3 Security

Security systems such as radio frequency identification (RFID) systems are commonly used to grant privileges (e.g. room and building access), and localization systems will be able to enhance these capabilities. If the correct person or people are in the correct

Specification	Value	Conditions
Accuracy	1.5 m	50% of estimates indoors
	5 m	50% of estimates outdoors
Range	>15 m	Indoors, through walls
	100 m	Outdoors, line of sight
Latency	< 5 s	Including data relay across network
Infrastructure Cost	Low	

Table 1.1 Summary of ranging specifications for typical indoor and outdoor sensor networks

rooms, privileges can be granted or revoked to ensure a secure environment for sensitive information. For example, certain prisoners in a prison may have access to certain resources, but this access may be denied if other prisoners are too close to the resource. Latency must be on the scale of a second, and accuracy must ensure correct room identification [11].

1.4.4 Summary of Specifications for ranging systems

Location accuracy, latency, range and infrastructure complexity are quite consistent across a broad spectrum of applications, and these requirements are shown in Table 1.1. For most networks a system with these specifications will provide a robust solution. Much higher accuracy may be required in some applications such as light switch replacement, but it is not all that common. Infrastructure points, or nodes, can vary in cost by orders of magnitude depending on the ranging method used, and reducing the cost of these points is important to a successful location aware wireless sensor network.

1.5 ORGANIZATION

Chapter 2 discusses the sources of error in RF time of flight ranging systems which are time synchronization, noise, quantization and environmental clutter. Chapter 3 introduces techniques compatible with low cost radios for reducing the impact of the error sources presented in Chapter 2. Chapter 4 presents a prototype platform and

implementation of a ranging system for wireless sensor networks. Chapter 5 presents the results from experiments carried out with this platform.

Chapter 2

Sources of Ranging Error

The achievable accuracy of ranging systems is limited by four primary factors which are noise, clock synchronization, sampling artifacts, and multipath channel effects. These factors introduce random, temporally and spatially varying errors into the range estimate resulting in limited accuracy. Time synchronization and frequency accuracy between the devices involved in the measurement can impact ranging system accuracy significantly because radio waves propagate so quickly that even minute timing errors can cause large measurement errors. Each effect can dominate the error under different circumstances, and a system must be designed so that the combination of these effects does not degrade accuracy beyond useful limits. Because the introduced errors are stochastic, the errors can never be eliminated, but it is possible that measurement techniques can be used to mitigate these effects. In this chapter, we discuss the various error sources and some methods for reducing these errors.

2.1 Noise

Noise and interference introduce unknown errors into measurements. The effect of white noise processes such as thermal and electronic noise is well understood and can be quantified. A range measurement degraded only by noise is limited in accuracy by the signal energy to noise ratio (SNR) at the receiver and the occupied bandwidth.

A ranging system suffers in a low SNR environment because the exact time of an event cannot be resolved precisely. In a simple example “edge detection” ranging

system, the ranging signal is a step function sent by the transmitter at $t = 0$ and the receiver measures the time of the rising edge it observes. When this signal is received, the edge time may be detected slightly early or slightly late due to noise added to the signal. For RF measurements radio waves move at the speed of light (3×10^8m/s) meaning that a distortion of just 10 ns results in 3 m of measurement error. The speed of this rising edge at the receiver is proportional to the bandwidth of the communications system, and wider bandwidth typically results in better performance. Because the noise amplitude increases as the square root of bandwidth and the signal transition speed increases linearly with bandwidth, a faster rising edge is more tolerant to noise. This qualitative understanding of how SNR and bandwidth affect the noise performance of ranging is useful, but a quantitative limit of ranging accuracy in a noisy environment is needed.

The mathematical expression that links SNR and bandwidth together to give a bound on ranging performance can be derived from the Cramér-Rao Lower Bound (CRB). The CRB can be calculated for any unbiased estimate of an unknown parameter. Ranging as a parameter estimation problem was widely studied in the context of radar and sonar applications, and the CRB has been derived under a variety of conditions [12]. For the prototype "edge detection" ranging system discussed above, the CRB can be used to calculate a lower bound for the variance of the estimate for the range, $\hat{r}$, as

$$\sigma_{\hat{r}}^2 \geq \frac{c^2}{(2\pi B)^2 \, E_s/N_0}\left(1 + \frac{1}{E_s/N_0}\right) \tag{2.1}$$

where $\sigma_{\hat{r}}^2$ is the variance of the range estimate, c is the speed of light, B is the occupied signal bandwidth in Hertz, and E_s/N_o is the signal energy to noise density ratio. The SNR is related to E_s/N_o in that

$$SNR = \frac{P_s}{P_n} = \frac{E_s}{N_0 t_s B} \tag{2.2}$$

where P_s is the signal power, P_n is the noise power, t_s is the signal duration during which the bandwidth, B, is occupied. The concepts of occupied bandwidth and signal duration are important as illustrated by our step function example. The maximum bandwidth of the signal is set by the transmitter filter, and increasing the receiver's filter bandwidth does not increase the bandwidth used by the signal. Similarly, t_s is not simply the length of time that the signal was observed at the receiver, but the length of time that the signal was observed when it was doing anything meaningful (such as changing in value). In the case of this step function, a small window of time contains nearly all of the useful information about the transition, and observing the signal for a longer time period contributes almost no additional information. In this example and in many common signals, the bandwidth and duration are tied together such that $t_s B \approx 1$. Therefore, the E_s/N_0 ratio is approximately equal to the SNR. By exchanging the locations of the factors in (2.2),

$$\frac{E_s}{N_0} = t_s B \cdot SNR \tag{2.3}$$

one advantage of having a $t_s B$ product greater than one becomes clear. Signals with this property would exhibit better noise performance at lower SNR values. One class of signals that exhibit this property are pseudorandom number sequences that result in long duration while retaining the same bandwidth as the constituent sub-symbols. These sub-symbols are called chips to differentiate them from bits (information) and symbols (collections of bits). Taking advantage of signals with $t_s B > 1$ improves noise performance, but it comes at the cost of increased signal processing. Often there is no other way to improve noise performance (i.e. the transmitter output power and

receiver sensitivity are fixed), and the signal processing cost is acceptable. For a fixed signal energy and noise density, increasing the bandwidth provides significant improvements in noise performance. This fact is one argument for increasing the bandwidth of RF based ranging systems, but the bandwidth required to achieve reasonable noise performance is not very large.

One common example can be found in GPS. The C/A (course acquisition or civilian) signal in GPS uses a pseudorandom number sequence modulated with binary phase shift keying (BPSK) at 1.023×10^6 chips/s. At a receiver on the ground, the observed SNR is typically -20 dB, the bandwidth occupied is about 2 MHz, and there are 1023 chips per symbol [13]. This is all the information required to determine the best case noise performance of GPS. First we calculate E_s/N_0 assuming a single 1023 chip sequence is observed through the application of (2.3):

$$\frac{E_s}{N_0} = t_s \cdot B \cdot SNR = \frac{1023}{1.023 \times 10^6} \cdot 2 \times 10^6 \cdot 10^{-2} = 20$$

Applying this result to (2.1)

$$\sigma^2_{\hat{r}_{GPS}} \geq \frac{(3 \times 10^8)^2}{(2\pi \cdot 2 \times 10^6)^2 \cdot 20}\left(1 + \frac{1}{20}\right) = (5.5m)^2.$$

This accuracy is close to what GPS routinely provides, but this range estimate is updated at 1kHz in the above calculation, and the typical user uses systems that update at less than 10 Hz. This can be used to reduce the variance by a factor of 100 (by increasing t_s by 100) resulting in $\sigma^2_{\hat{r}_{GPS}} \geq (0.6\text{m})^2$. GPS users are accustomed to accuracy of 5 m (80% of trials) in open, flat terrain suggesting that the noise limit is not obtained or that other factors are reducing accuracy. In this case, approaching the CRB is possible because of the high value of E_s/N_0 and the signal design, but random atmospheric effects contribute the majority of the remaining error. The P (precise or military) GPS

signal is broadcast at two different carrier frequencies so that these atmospheric effects can be estimated and removed greatly enhancing accuracy [13]. It is also worth noting that the $1 + E_s/N_0$ term contributes very little to the CRB, and it is commonly ignored for $E_s/N_0 \gg 1$.

GPS provides a good reference for looking at other ranging systems because it is familiar and has some characteristics in common with communications systems, but it has significant differences as well. In typical wireless communications systems, the distances traveled are much less, and atmospheric effects are not significant. In addition, narrowband communication systems have high SNR such that, when coupled with processing gain, very high values of E_s/N_0 result. These high values for E_s/N_0 allow the CRB to be nearly achieved in many systems, but the CRB is not a tight bound at low E_s/N_0 [12]. If the desired error variance is not achievable directly, averages of multiple measurements will yield improved results. Both GPS and communication systems must contend with multipath propagation in the channel, and this multipath interference negatively impacts accuracy. GPS occupies a 2 MHz bandwidth which is the same as the common IEEE 802.15.4 radios used in wireless sensor networks, but GPS signals are broadcast at a single carrier frequency. This combination of a narrow bandwidth and single carrier frequency makes GPS particularly susceptible to large multipath induced errors. WSN radios are usually frequency agile, and information from different frequencies can be used to improve ranging performance in these difficult environments. [14].

The CRB can also be improved through the use of additional bandwidth. Ultra-wideband (UWB) technologies are being developed partially to provide accurate ranging capability to wireless systems. An UWB signal is defined to be a signal that

either uses at least 500 MHz or that occupies as much bandwidth as half of the signal's center frequency. The use of 500 MHz of bandwidth and an E_s/N_0 of -10dB yield a CRB of

$$\sigma_{\hat{r}}^2 \leq \frac{(3 \times 10^8)^2 \left(1 + \frac{1}{0.1}\right)}{(2\pi \cdot 500 \times 10^6)^2 \cdot 0.1} = (1m)^2.$$

Although the CRB may not be achievable at this low value for E_s/N_0, small bounds are possible. This promise, along with superior performance in multipath environments (to be discussed later), has driven much interest in UWB for extremely accurate location systems.

This work considers the low power narrowband radios already in widespread use even though UWB is the primary focus of most research on wireless ranging. UWB radio transmitters are simple to design and are very low power making them attractive for low power devices. UWB receivers, on the other hand, are very complex and power hungry and/or have very poor performance in real environments. The primary limiting factor for UWB receivers is linearity in the presence of narrowband interference. Low power UWB receivers like those that will comply with the new IEEE 802.15.4a standard are designed to detect energy in the channel, and energy detection schemes are not robust to narrowband interference [15]. This work considers narrowband radios because current low power radios perform very well in real environments, and narrowband radios will continue to play the leading role in reliable, low power wireless connectivity for the foreseeable future. As a result, enabling accurate ranging in narrowband systems is important because narrowband systems will continue proliferate in wireless connectivity space.

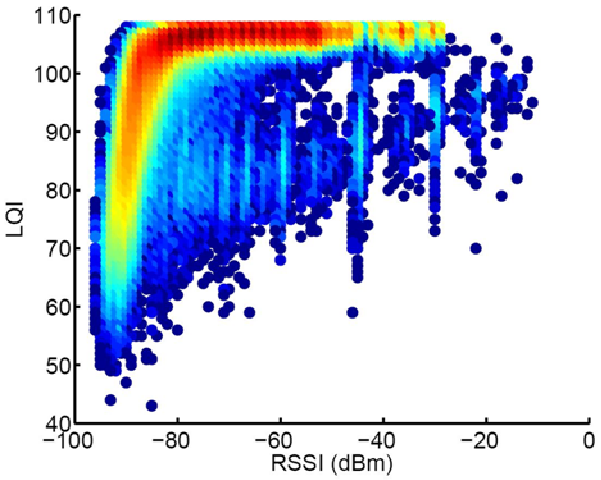


Figure 2.1 The number of links at a given RSSI and LQI (Link Quality Indicator output, a measure of SNR) is shown in this density plot. Red points indicate numbers on the order of 10^4 while dark blue points indicate numbers on the order of 10^0.

Both bandwidth and E_s/N_0 play significant roles in determining noise limited performance, and it is important to understand typical conditions in wireless local area and sensor network environments. Figure 2.1 shows a density plot of the number of packets with a given RSSI and SNR (shown as LQI, an arbitrary unit for SNR) for some 9 million packets exchanged over a Dust Networks test network in a factory [16]. It is apparent that many of the paths are at high SNR, and typical baseband SNRs (P_s/P_n) range from 8 dB to 28 dB. About 85% of the links have SNR above 10dB, and 50% of the links have SNR above 20dB [17, 18]. Signals with t_sB products ranging from 10 to over 1000 (10 dB to 30 dB) are commonly used enabling very large E_s/N_0 in communication systems. Figure 2.2 shows the CRB as a function of bandwidth for E_s/N_0 of 10 dB and 26 dB. It is interesting to note that noise alone does not prevent 1 m accuracy for bandwidths of a few megahertz or more.

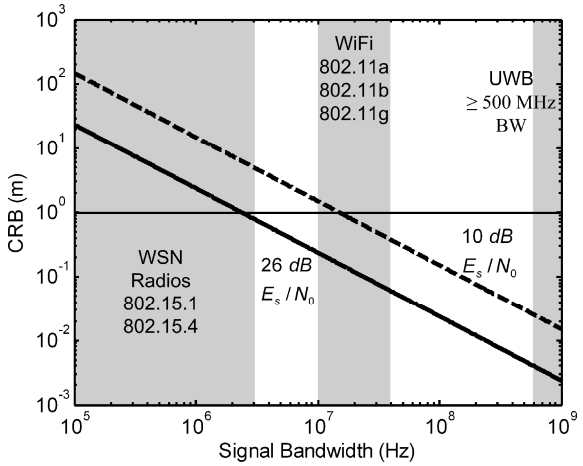


Figure 2.2 Cramér Rao Lower Bound (CRB) as a function of bandwidth for 10dB and 26dB E_s/N_0. Common radio standards used in wireless sensor networks such as IEEE 802.15.1 (Bluetooth), IEEE 802.15.4 (Zigbee and others), and wireless LAN (802.11a/b/g) are shown. The CRB predicts that ultra-wideband (UWB) radios will have excellent noise performance, but the CRB is not a tight bound for the low SNRs (not shown in plot) observed in UWB systems. It is expected that UWB ranging will have noise limited accuracy of better than $1m_{RMS}$ in most practical cases. Even a few megahertz of bandwidth can enable the 1.5 m accuracy required for most applications.

2.2 CLOCK SYNCHRONIZATION

Time of flight measurement systems must be able to estimate the time of transmission and arrival using a common time base for accurate measurements. When two wireless devices, A and B, perform range estimation, the most straightforward method is for A to send a signal at $t = 0$ and for B to start a timer at $t = 0$ and stop it when it receives the signal sent by A. The value of the timer at B is equal to the TOF. This method is shown in Figure 2.3a. If the clocks are not perfectly synchronized, however, and B's notion of $t = 0$ is offset in time from A's, then this offset, Δt, directly adds a bias to the measurement. Time synchronized wireless networks are typically synchronized on the order of one bit period, T_{bit}. In typical systems, T_{bit} ranges from 0.1 μs and 1 μs resulting in errors of between 30 m and 300 m.

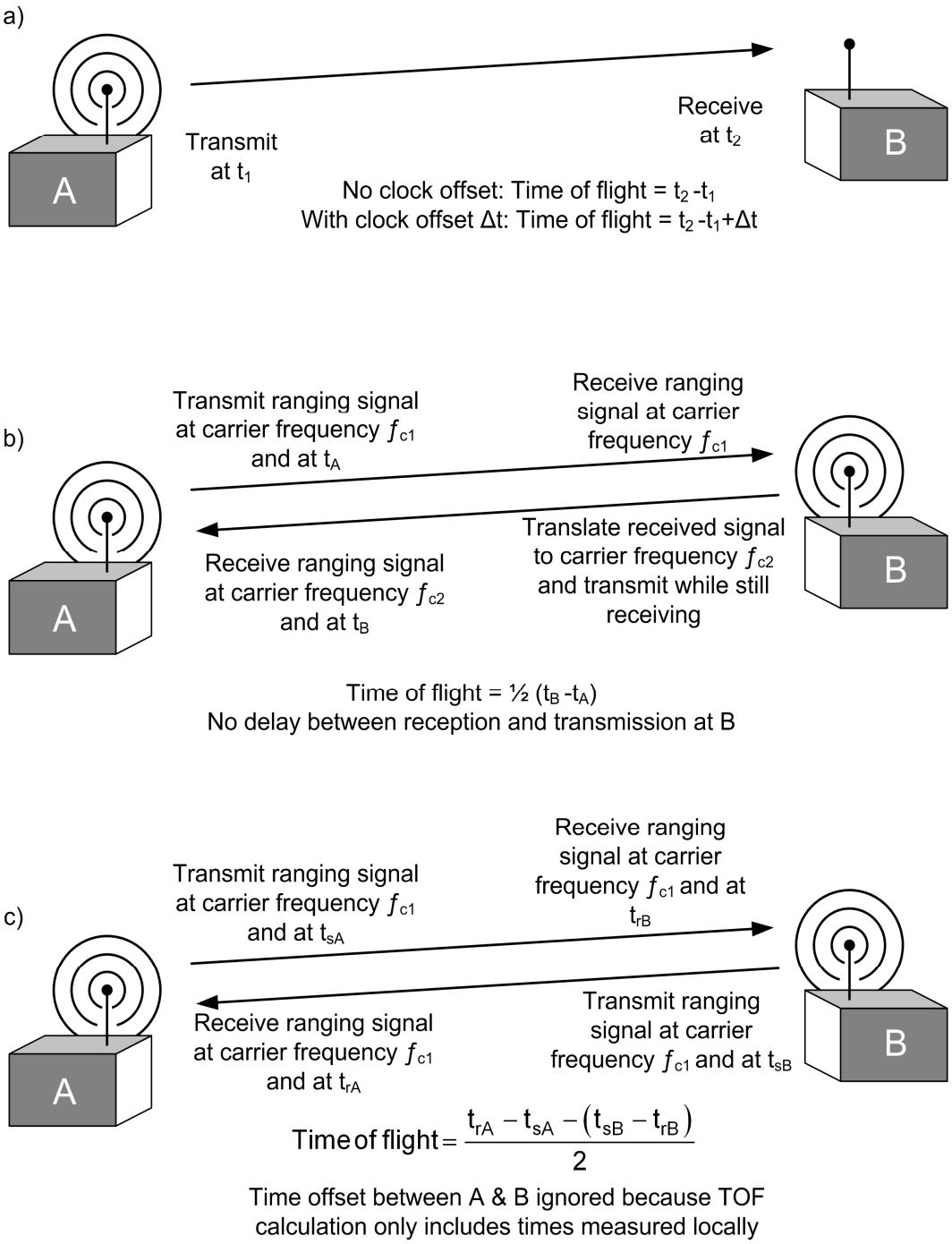


Figure 2.3 Three methods of performing time of flight ranging measurements: a) time of arrival which is susceptible to clock offset Δt; b) full duplex two way ranging; c) half duplex two way ranging called two way time transfer.

High power and expensive systems can achieve time synchronization of better than 10 ns or 3 m. GPS satellite time synchronization is maintained to within 10 ns, and some terrestrial systems are synchronized to better than 1 ns.

If A and B have full-duplex radios, that is, they can transmit and receive at the same time, then a two way or round trip measurement can be made. A sends a signal to B at a center frequency f_{c1} and B translates this signal to a different carrier frequency f_{c2} and retransmits that signal in real time. The signal is received back at A at f_{c2} such that A can compare the signal it is receiving from B to the signal it is sending to B. By measuring the delay between these two signals, the round trip TOF, $\hat{\tau}_{RT}$, is estimated, and the range estimate is $c \cdot \hat{\tau}_{RT}/2$. This method is shown in Figure 2.3b. Full duplex two way ranging has been used successfully since its first use in the Second World War, and it is generally deployed on top of standard radar systems for tracking civilian aircraft. The airplane transponder mixes the incoming radar waveform to a new carrier frequency and transmits the incoming signal back. Additional information is added providing more detailed information on location, heading and velocity. Aircraft transponder accuracy is generally reported to be better than 100 m [18].

Most WSN nodes do not have full-duplex radios because they are more complicated and expensive than half-duplex transceivers. Many other wireless systems are half duplex as well (e.g. wireless LAN and GSM), and the round trip method can be adapted for half-duplex systems. A round trip method known as two way time transfer (TWTT) has been developed to improve time synchronization between wireless base stations after the first communications satellites were launched, and it provides both range estimation and improved time synchronization capability [19]. This method, shown in Figure 2.3c, allows the time offset between A and B to be ignored. Both A and

B are responsible for measuring a time delay accurately using a local clock. Node A must measure the time that it takes for the signal it sends to return to it, and B must measure the time that the signal spends at B accurately. If the time A sends the signal is t_{sA}, the time B receives the signal from A is t_{rB}, the time B replies to A is t_{sB}, the time A receives the signal is back from B is t_{rA} such that $t_{sA} < t_{rB} < t_{sB} < t_{rA}$ then A measures $t_A = t_{rA} - t_{sA}$ and B measures $t_B = t_{sB} - t_{rB}$. By combining these two measurements together both the time of flight ($\hat{\tau}$) and clock offset ($\Delta\hat{t}$) can be estimated.

$$\Delta\hat{t} = \frac{1}{2}(t_A + t_B) \tag{2.4}$$

$$\hat{\tau} = \frac{1}{2}(t_A - t_B) \tag{2.5}$$

This or related methods are used with less accurate hardware to provide the rough time synchronization common in wireless systems.

The noise performance of TWTT measurements is easily found when considering equations 2.1 and 2.5. A TWTT measurement is simply the average of two of the measurements considered in (2.1), and, therefore, the resulting noise performance for high values of E_s/N_0 can be found.

$$\sigma_{\hat{r}}^2 \geq \frac{c^2}{2(2\pi B)^2\, E_s/N_0}.$$

From this result, we can see that TWTT has a slight noise benefit over one way ranging at the expense of roughly twice the energy consumption.

One problem with two way ranging is that the measurement takes place over a relatively long period of time such that if the reference frequencies at the two nodes are not identical, an unknown bias can be added to the signal. In WSN nodes, inexpensive

crystals are used where the frequency spread from crystal to crystal may be 100ppm or more across commercial temperature ranges. This clock frequency offset error (also called clock drift) must be mitigated in some fashion [14]. Consider a system where a ranging signal is sent for 100μs, the time to switch between transmit and receive is 200μs, and then the signal is received for 100μs. Over this 400μs time, a clock frequency mismatch of just 10ppm would result in about 4ns or 1m of estimation error. The clock frequency offset can be measured, and then the clock frequency can either be corrected to match within bounds or the resulting error can be calculated and subtracted from the estimate later. Many methods have been used to measure frequency offsets in wireless systems, and we summarize one simple method here. This method is to run a counter over a long period of time to measure the offset. One node sends a start packet to the second node and starts a local timer, and the second node starts a local timer when it receives this packet. After waiting a sufficiently long time, the timer at the first node expires, and it sends a stop packet. The second node receives this stop packet, stops its timer, and compares the value left on the timer to the expected value (zero if the counter is counting down). This difference is a measure of the clock offset. The minimum time between packets, T_{wait} can be calculated as follows:

$$T_{wait} \geq \frac{1}{\Delta f_{xo}} \tag{2.6}$$

where Δ is the required matching between local frequency references, and $f_{xo}$is the the local reference frequency. For a 20MHz crystal and a system requiring 10ppm accuracy, T_{wait} must be great than 5ms. This process is rather long but very simple, and other methods trade complexity for time savings.

2.3 SAMPLING ARTIFACTS

Modern ranging systems estimate the time of flight by sampling the incoming signal and estimating its time of arrival based on these samples. It is often asseted that ranging accuracy is limited to c/f_s where f_s is the receiver sampling rate [20]. This limit is known as range binning, and it can impact resolution if steps are not taken to mitigate its impact. A common implementation is to estimate the time of arrival using a matched filter that is sampled at up to twice the signal bandwidth resulting in time resolution of $1/2B$. This sampling adds error to the estimate because the estimate space is divided up into range bins that are $c/2B$ wide. The error associated with this process is uniformly distributed inside the range bin. By using the variance of the uniform distribution, the impact of sampling can be calculated [21].

$$\sigma_{sample}^2 = \frac{1}{12 \cdot f_{sample}^2} \tag{2.7}$$

In the case of the GPS example, with sampling at $1/2B$ the variance due to sampling can be calculated.

$$\sigma_{sample}^2 = \frac{1}{12 \cdot (4 \times 10^6)^2} = (72ns)^2$$

This results in a range resolution of 22 m. In GPS, this coarse estimate is filtered (averaged) to improve the resolution, and a feedback loop can be used to null out the sampling error while the receiver tracks the satellites [13]. Using just averaging, over 450 measurements are required to achieve a variance of $(1m)^2$. These methods are not realistic for many WSN applications where extremely low power consumption and therefore duty cycle is required. An accurate range estimate must be made in a short period of time. To reduce the sampling error, the signal can be over sampled. Figure 2.4 shows the CRB for a 2 MHz bandwidth signal with E_s/N_0 of 26 dB, the standard

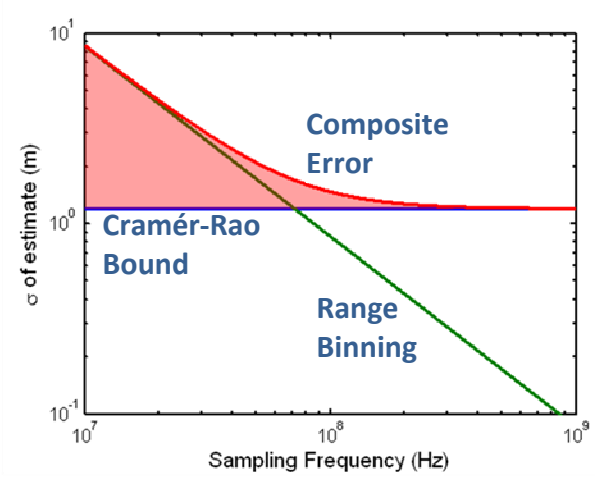


Figure 2.4 A comparison of range binning due to sampling error and the Cramér-Rao bound on noise limited ranging for a 2 MHz bandwidth with a E_s/N_0 of 26 dB. The sampling rate required is much higher than required by sampling theory to achieve noise limited resolution. The shaded region represents the accuracy sacrificed due to range binning compared to a Cramér-Rao bound limited system.

deviation of the range error due to sampling, and the combined effect of both error sources as a function of sampling frequency. This plot shows that with a 2 MHz bandwidth, the required sampling rate to ensure that the error is not dominated by sampling is over 70 MHz. It is clear that one must sample very fast to have the error dominated by the CRB rather than sampling. As the CRB improves due to increased bandwidth, the sampling speed required remains higher than twice the signal bandwidth down to E_s/N_0 of about 3 dB.

If the signal is sampled above Nyquist ($f_{sample} > 2B$), then the entire information content of the signal is captured in the sampling process [22]. Therefore, it should be possible to extract better time resolution than σ_{sample}. In Figure 2.5 a signal is shown along with dots representing the samples of that signal that is band limited to a 2 MHz bandwidth. This signal is sampled at 10 MSps which is above the Nyquist rate of 4 MSps, but the sample rate still is far too low to achieve the CRB. The range bins are 100 ns (30 m) wide in this case where as the CRB from Figure 2.4 is only 3.5 ns (1.1 m) demonstrating a dramatic resolution reduction. Looking at the time of the zero crossing,

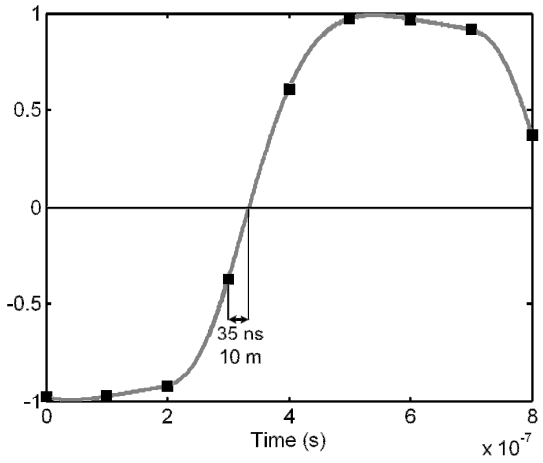


Figure 2.5 An above Nyquist sampled waveform is shown with the sample points marked in an example of sample based range binning. An interpolation between points enables time resolution of the zero crossing far better than 1/B and $1/f_s$ reducing the size of the range bins significantly.

it is clear that even a linear interpolation between the two adjacent samples would improve the estimate of that zero crossing location significantly.

A major challenge is that current two-way ranging methods need to perform time of arrival estimation in real time (at node B where the signal reply occurs). In practice the algorithm for estimating time of arrival is more complicated than just estimating a zero crossing time, and estimating the time of arrival is time consuming and processor intensive. The time of arrival estimation algorithm used in the system proposed in this work takes more than 1.6ms to compute, but the typical transmit to receive mode switching time is less than 200 μs. Adding this time increases the required frequency matching between the devices significantly greatly adding complexity and energy costs to the ranging operation.

2.4 Multipath Channel Effects

When a ranging system has been well designed, it often still fails to achieve the expected performance because the measurement is not taken in free space. In real environments the RF signals bounce off objects in the environment causing the signal to arrive at the

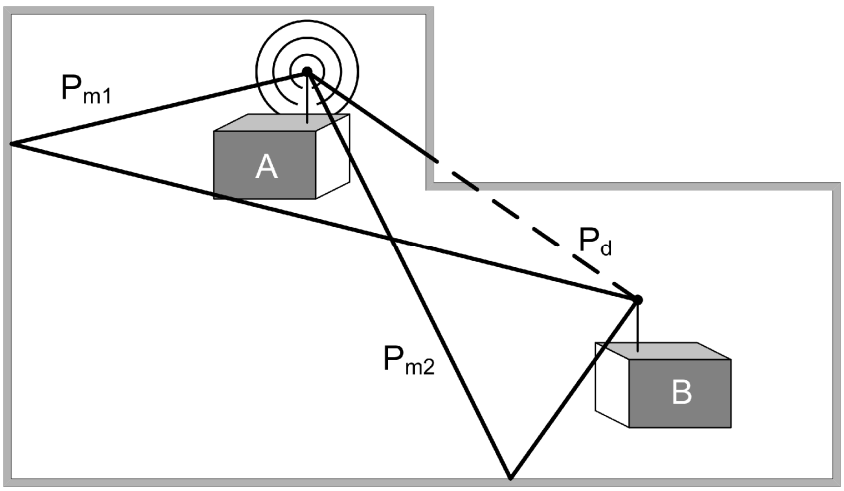


Figure 2.6 A multipath environment that exhibits a common condition. The direct path (P_d) which is to be estimated for ranging is obstructed and heavily attenuated while the reflected paths (P_{m1}, P_{m2}) have much higher signal power.

receiving antenna through multiple paths as shown in Figure 2.6. In this figure, the direct path is obstructed by walls, but the other paths are not. This is common indoors, and it is possible that the non-direct paths have higher power than the direct path [23]. The communication environment is called the channel, and multipath channels are not only specific to the type of environment (office building, residential or outdoors) but to the specific geometry of the transmitter and receiver in that environment. In the most general case, the channel impulse response can be modeled as a series of complex delta functions in time.

$$h_c(t) = \sum_{i=0}^{N} A_i \delta(t - \tau_i) e^{j\phi_i}$$

where A_i, τ_i and ϕ_i are the amplitude, time and phase delay of the ith path. The amplitude, time delay, and phase are all random parameters, and a variety of distributions are commonly applied to them [24]. The transmitted signal can be represented using the phasor notation of the RF signal.

$$m(t) = \mathrm{Re}\{ e^{j(\omega t + \theta(t))} \}$$

In $m(t)$ the time dependent phase term can represent frequency of phase modulation, and the signals considered have constant amplitude that can arbitrarily be set to unity. The resulting received signal is the convolution of the transmitted signal and the channel response with complex additive white noise.

$$s(t) = m(t) * h_c(t) + \tilde{n}(t)$$

The noise term, $\tilde{n}$, will be ignored in this analysis as it does not impact multipath performance. If $h_c(t)$ consists of just two paths, we can easily write the entire received signal, $s(t)$.

$$s(t) = \mathrm{Re}\left\{A_0\, e^{j\left(\omega(t-\tau_0)+\theta(t-\tau_0)\right)} e^{j\phi_0} + A_1\, e^{j\left(\omega(t-\tau_1)+\theta(t-\tau_1)\right)} e^{j\phi_1}\right\}$$

The terms with a subscript 0 are from the direct path (assuming one is present), and the terms with subscript 1 are due to multipath propagation. There is an additional phase term that depends on carrier frequency, and it can be pulled out of the main exponential.

$$s(t) = \mathrm{Re}\left\{A_0\, e^{j\left(\omega t+\theta(t-\tau_0)\right)} e^{-\omega\tau_0} e^{j\phi_0} + A_1\, e^{j\left(\omega t+\theta(t-\tau_1)\right)} e^{-\omega\tau_1} e^{j\phi_1}\right\}$$

This term is often combined with the ϕ_i term and modeled as a random parameter in communication systems, but it is important to note that this term causes the channel to be frequency dependent in both amplitude and phase response.

The channel is often time varying resulting in a multipath environment that changes from one time to another. For narrowband radios like those common in WSNs, moving one transceiver by just a fraction of a wavelength ($\lambda = 12$ cm at 2.4 GHz) will cause the receiver to see what looks like an entirely new multipath environment because the paths will interfere constructively or destructively differently. The path length change is referenced to the wavelength of the RF making these small changes have large effects. The speed that the channel changes depends on how quickly objects

are moving in that environment. Slower objects result in slower changes to the channel. This typically means that indoor channels change more slowly than outdoor channels, and the time it takes for the channel to change significantly is called the coherence time, t_c, of the channel. The value of t_c is roughly $c/2fv$ where c is the speed of light, f is the carrier frequency, and v is speed of the fastest moving object in the environment. Recall that the wavelength of radio waves, λ, is c/f, and a more intuitive form of t_c is $\lambda/2v$ where it is clear that the time it takes to move a half wavelength corresponds to the coherence time [24]. In indoor environments, people and things move rather slowly. People walk at 2 m/s, and some objects in industrial settings may move at up to 5 m/s. The coherence times at 2.4 GHz for these examples are 31 ms and 13 ms respectively.

A series of measurements that take less than t_c to complete can be used together as if the channel were time invariant over those measurements. This fact is useful when attempting to reduce the impact of multipath propagation because multiple measurements taken at different frequencies can be used together. Because this interference effect is closely tied to the wavelength, changing carrier frequency even by 1% or less can dramatically affect the apparent multipath environment in narrowband systems. This can be observed by considering the received signal strength (RSS) profile across carrier frequency in an indoor environment as shown in Figure 2.7 [1]. At some carrier frequencies, the signal experiences deconstructive interference (referred to as fading), while at others it has much higher signal strength due to constructive interference. Without knowing the channel characteristics, knowledge of the RSS at one frequency tells you nothing about the RSS at a nearby or distant frequency. Wider bandwidth signals suffer less from this effect, and the bandwidth required to combat this effect is related to the time difference between the first and last significant path

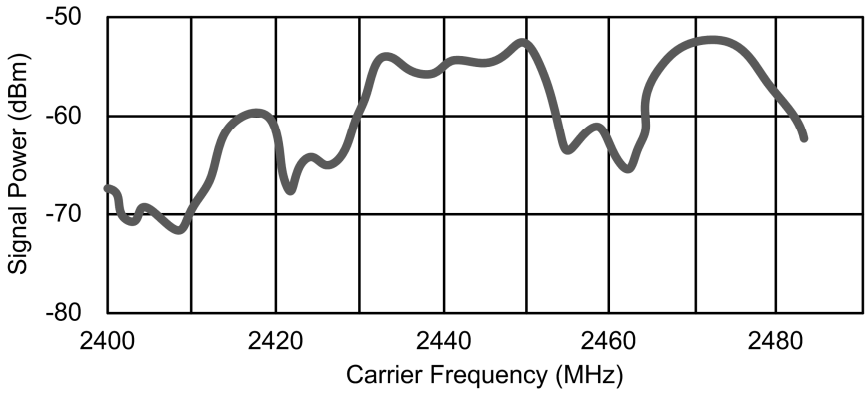


Figure 2.7 Received signal strength verses frequency measured in a line of sight multipath channel with a 2MHz RF bandwidth. The significant changes in signal strength show that changing carrier frequency changes the apparent multipath environment significantly. Adapted from [1]

arrivals known as the delay spread, t_d. The coherence bandwidth, W_c, is approximately $1/2\pi t_d$ and it is the bandwidth over which the channel can be considered to be flat (either in deep fade or not, for example). If the bandwidth, B, is much larger than W_c the signal does not depend on carrier frequency to the same extent as a signal with a bandwidth less than W_c [24]. Typical delay spreads for indoor channels are between 10 ns and 100 ns yielding coherence bandwidths between 1 MHz and 20 MHz. Outdoors, the delay spread can be up to microseconds, significantly reducing W_c. In ranging systems, the inter-path delay, $t_{\Delta p}$, is more important than the delay spread, however, because short inter-path delays can significantly impact ranging accuracy. Indoors, inter-path delays of 5ns to 10ns are very common and must be resolved if accuracy is to be better than $c \cdot t_{\Delta p}$ [12].

In a multipath environment, the receiver must somehow choose or estimate the direct path and ignore the other paths. If a receiver can estimate when only the first path arrives, then this will be the shortest distance and thus the desired estimate. If the system is not able to resolve the individual paths, then the estimate is blurred by the multipath effects resulting in estimation error. In this case, if the receiver has an

estimate of the channel impulse response, it can estimate the bias caused by the multipath channel and subtract the bias from its estimate. This leads to two classes of multipath mitigation methods: 1) resolving the direct path through increased bandwidth, or 2) using channel information to improve a narrowband range estimate.

In the first case, the ability to resolve the response of the multipath channel is directly linked to the bandwidth of the signal. Inter-path delays, $t_{\Delta p}$, separated by more than $1/B$ in time are resolvable and paths separated by less are generally not. To resolve paths that are separated by 1m or more, a bandwidth of at least 300 MHz is required, showing a significant advantage of UWB systems. Using bandwidths in excess of 500 MHz enables accuracy better than 1m in many cases, but this accuracy is not always achieved [25]. When the direct path is too weak compared to other paths, a secondary path will be chosen to estimate the range resulting in an over-estimate. In indoor environments, 10% to 20% of all measurements will fall into this category, but some environments are worse and a direct path is rarely available. Note this is different than a situation in which a line of sight, or unobstructed, path is available. Although line of sight paths can be common with good geometries indoors, most indoor channels will have a few strong paths spread across a few tens of nanoseconds [23]. UWB systems have been demonstrated to provide ranging accuracy better than 1m [26], but few demonstrated systems exist. The UWB systems that do exist have not approached the low power capabilities of narrowband radios.

A second method for attempting to mitigate the impact of multipath interference is through super-resolution ranging methods where a larger bandwidth is synthesized from 1 or more narrowband measurements. A super resolution algorithm is one that attempts to provide range resolution that is better than $1/B$ [27]. Super-resolution

methods come in two flavors: A) methods that coherently combine multiple measurements across different carrier frequencies, B) methods that estimate the channel characteristics at a single carrier frequency.

Coherent combining of multiple measurements, option A, is practical in orthogonal frequency division multiplexing (OFDM) systems where coherent measurements can be taken simultaneously at many center frequencies. In this case, the frequency response (magnitude and phase) can be measured directly by measuring the carrier pilot signals. In systems that must frequency hop to measure on multiple channels, this method is extremely difficult to implement. Coherent measurements are challenging because large phase rotation errors accumulate in very short times (such as the time to change channels), and estimating these errors is often cost prohibitive. Although in the OFDM case it would seem the estimate would be limited in time resolution to $1/B$ the multiple signal classification (MUSIC) algorithm commonly used in this method provides resolution better than 1/4B in many cases [28]. In the case of IEEE 802.15.4 with a 2 MHz bandwidth, the achievable resolution is approximately $100ns$ or $30m$. This is still insufficient to achieve the accuracy required for the applications of interest.

Option B relies on estimating the impact of the multipath environment on the range estimate from a single measurement. This method, implemented using the matrix-pencil algorithm [29], is used when the signal bandwidth is too small to sufficiently resolve the multipath environment and there is sufficient E_s/N_0 to resolve meaningful channel information. It is somewhat analogous to channel equalization, and both ranging and equalization can utilize the same channel estimate. To estimate the channel impulse response, a known, modulated signal consisting of a sequence of chips

is sent through the channel [30]. Recall that the inter-path delay is a few nanoseconds compared to the chip duration of 100's of ns to μs, and the chip width in previous methods must be shorter in time than the features to be resolved. If the signal sent is x, the channel impulse response is h, and the received signal is y, then

$$y = x * h + \tilde{n}$$

Where $*$ denotes convolution, and $\tilde{n}$ is complex noise. This can be rewritten in the frequency domain.

$$Y(\omega) = X(\omega)H(\omega) + N(\omega)$$

If the signal to noise ratio is large, and the spectrum of the transmitted signal (including the transmitter frequency response) is known, then $\widehat{H}(\omega)$ can be approximated.

$$\widehat{H}(\omega) = \frac{Y(\omega)}{X(\omega)} + \frac{N(\omega)}{X(\omega)} \approx \frac{Y(\omega)}{X(\omega)}$$

This approximation is only valid in sufficiently high SNRs, and noise causes significant estimation errors. $Y(\omega)$ is calculated by taking the FFT of the received signal, and $X(\omega)$ is a system parameter known a priori. Once $\widehat{H}(\omega)$ has been estimated, $\hat{h}(t)$ can be estimated. The inverse Fourier transform will solve this problem, but a number of substantially more complicated algorithms exist that provide better time resolution [29, 31]. These algorithms may achieve time resolution that is four times better than the Fourier transform method when the SNR is high enough. In 802.11b systems, it is believed that these methods may be able to provide reasonable performance although this has not been demonstrated. In 802.15.4 systems, the performance is insufficient. The computational complexity of the algorithm is significant and is outside the scope of algorithms to be implemented on embedded processors used in wireless sensor networks, although some have proposed it may be possible to port similar algorithms to

low cost 802.11b devices [32]. The estimated computational time of the algorithm in [32] on an MSP430 class microcontroller at 25 MHz is 12s (30 million operations). In dedicated silicon with wider and more accelerated multiplication and division functions, it is expected these computation could be done in many hundreds of milliseconds. The time (and resulting energy) cost of such an algorithm suggests it is outside the scope of most wireless sensor network applications.

2.5 Summary of Performance Limits

In WSNs, the devices are resource and energy limited, and efforts should be made to reduce the time the radio is active and reduce the amount of signal processing while preserving performance. The above discussions show that signal bandwidth is a system parameter of high importance. Increasing signal bandwidth can improve noise and multipath performance linearly with bandwidth. The bandwidth required to achieve very fine resolution in a Gaussian white noise environment is far smaller than that required to achieve equivalent resolution in a typical indoor multipath environment, and the techniques to improve multipath performance are far more intensive than those to combat noise. Many measurements in indoor environments will not have a resolvable direct path using any method, and the resulting range estimate will be highly inaccurate. Localization algorithms must deal gracefully with range measurements that are widely inaccurate some of the time. Methods to deal with other error sources such as synchronization and sampling exist and should be applied to minimize energy while maximizing performance. Although UWB systems are sure to provide fine range resolution, the energy cost of data communication over an UWB radio remains very high compared to narrowband radios. Therefore, ranging methods that use small

bandwidths are critical to many low power wireless networks, and methods to improve range accuracy given fixed, small bandwidths are an unsolved problem.

Chapter 3

Ranging Error Mitigation Techniques

Accurate range measurements are the key to accurate localization in local area networks, and Chapter 2 introduced the sources of error that make ranging a challenging problem. In this chapter, we discuss a combination of new methods that, when combined together, provide meter level accuracy at significantly reduced complexity and without the need for multiple coherent channel measurements. In this chapter, each proposed technique will be presented along with the error sources it attempts to combat.

3.1 Code Modulus Synchronization

Code modulus synchronization is a two way ranging method that has better noise performance than two way time transfer while being optimized for low computational overhead and Nyquist sampling to avoid range binning. Code modulus synchronization is used to mitigate the effects of noise, clock synchronization, and sampling artifacts. This section contains a description of the method and analysis of its noise performance compared to other published methods [33].

3.1.1 Full Duplex Two Way Ranging

The inspiration for code modulus synchronization is a full duplex ranging operation. In a full duplex ranging operation, the time of arrival is calculated just once at the signal source, and the 2nd participating node only reflects the signal without further processing. The method with a simple baseband signal is shown in Figure 3.1. Signals

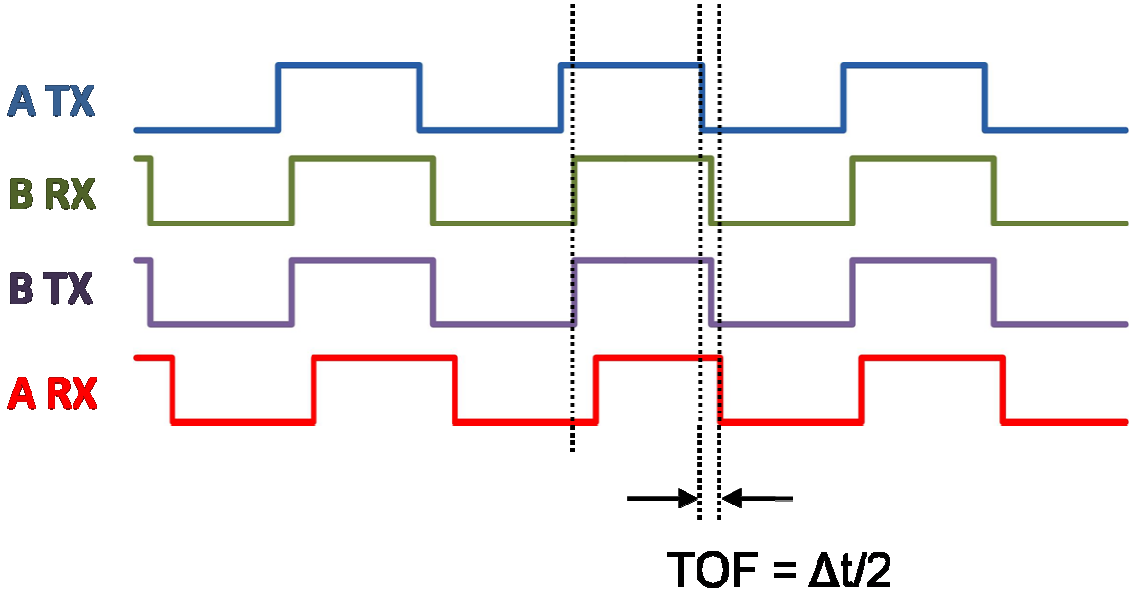


Figure 3.1 Baseband signals for full duplex two way ranging. The transmitter signals are presented on lines labeled TX, and the receiver signals are presented on lines labeled RX. A signal (a pulse train in this case) is sent from A to B. B receives the signal, mixes (carrier frequency translates) it to a new frequency and transmits it back with no delay. A receives the signal while it is still transmitting, and compares the two to determine the time of flight. This method has been widely used in since the 1940s.

with large Bt_s products are easily incorporated in this system enabling good performance in noisy environments. There are two primary advantages of determining the time of arrival a single time (as compared to the two times in two-way time transfer presented in 2.2). The first advantage is that instead of averaging two estimates for an $1/\sqrt{2}$ improvement in noise performance, the single estimate is divided by two to yield the time of flight. This results in a noise improvement of $1/\sqrt{2}$ over two way time transfer. The second advantage is that because the signal is only analyzed once, the reference and received signals can be analyzed offline if they are digitized and stored during the online communication portion of the measurement. The advantage of this is that range binning can be prevented by sampling the signal above the Nyquist rate and interpolating the signal to achieve the noise limited resolution. The problem with this

method is that the radios used in today's local area networks are half duplex, and it is impossible to implement this scheme.

3.1.2 Code Modulus Synchronization Algorithm

Code modulus synchronization emulates a full duplex ranging system, but half duplex radios such as those used in WSNs are used so the delay between reception and retransmission must be managed carefully. Code modulus synchronization uses a periodic signal (such as a square wave or a pseudorandom code) modulating an RF carrier as the ranging signal so that large Bt_s is possible through processing gain. Figure 3.2 shows the basic operation of the CMS using a square wave baseband signal. The first node, C, generates a local baseband ranging signal shown on the top line (C REF/TX) of Figure 3.2. This code is used to modulate the carrier and, in the shaded region, is transmitted to the second node, D. D has a local clock with the same period as at C, but the phase of the clocks are offset. As a result, D knows the length of the incoming code, but it does not know the phase offset in the clocks. D samples and demodulates this signal, and exactly one circularly shifted copy of the code is stored in memory (shown on line 2, D RX, of Figure 3.2 in the shaded region). At this point, D has a local copy of the code that is circularly shifted due to the clock phase offsets between C and D, and this reference code is shown on line 3 (D REF/TX) of Figure 3.2. After C has sent the code and D has received the code, the transceivers switch states, and D is now the source of the code. Node D transmits two copies of the circularly shifted code it received back to C, and this transmission is shown in the shaded box over line 2 (D RX) of Figure 3.2. Node D receives the signal and records it synchronized to its local reference shown on line 1 (C REF/TX). Because of the roundtrip nature of the system, the circular shift that occurred going from C to D is exactly undone going from D to C.

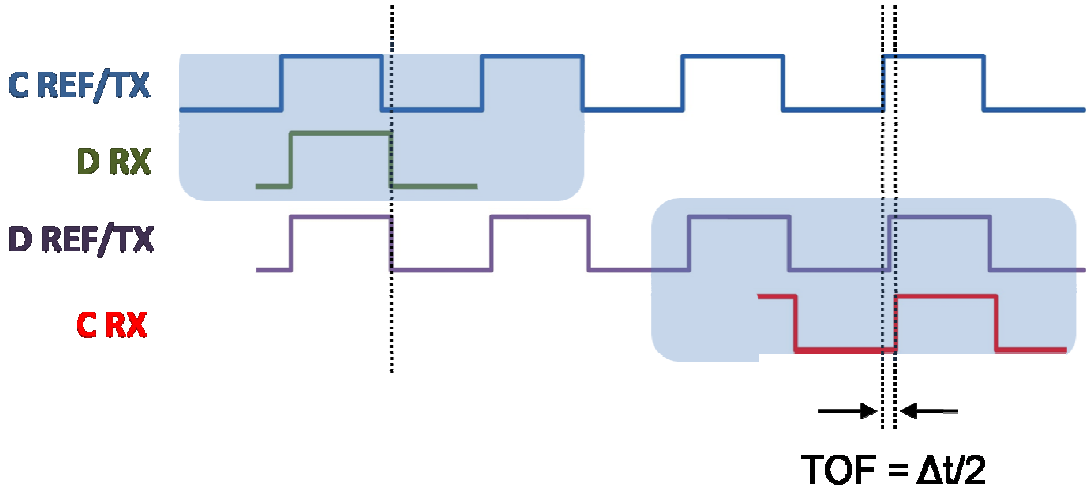


Figure 3.2 Baseband signals for code modulus synchronization, a half duplex two-way ranging method. The shaded regions represent when a signal is being transmitted. This figure and Figure 3.1 are largely equivalent because code modulus synchronization is an adaptation of Figure 3.1 for half-duplex radios.

After C has received the code, the transceivers are shut off, and all of the real time processing is completed. Node C then computes the cross correlation between the code it recorded and the code that it sent, and the measured code offset is the time of flight. Because this system relies on sampling the signal above Nyquist, the received code can be interpolated to improve resolution up to the noise limit of the system. The correlation and code offset estimation are not done in real time enabling the computation to be done at any time using any method the user desires. This system can approach the CRB in a single measurement, substantially improving over other two-way ranging methods. Code modulus synchronization is an adaptation of full-duplex ranging for half-duplex radios, and these two methods provide equivalent performance.

It is possible to send multiple copies of the code in order to increase E_s/N_0. The receiving system can accumulate (or average) multiple copies of the code in order to improve SNR, but they are all exactly one copy of the code that is circularly shifted in exactly the same way as the other received copies. This averaging of multiple copies is

important for achieving good noise performance, but it does not change the system's ability to resolve the time of flight accurately.

3.1.3 Code Length Considerations

The length of the code is chosen such that the time duration of the code is larger than the maximum range measurement of interest. If the code length is too short, range ambiguity can occur. For example, if the code length was T_c, the maximum non-ambiguous range would be $c \cdot \frac{T_c}{2}$ due to the roundtrip nature of the ranging operation.

3.1.4 Noise Analysis

In two-way time transfer (TWTT, see Figure 2.3c), the time of arrival must be determined at both nodes involved in the range estimation, but in CMS only one node performs this calculation. Therefore while CMS reduces the required real time processing enabling better sampling performance, the full processing gain of the system is not realized at the second node in CMS. This causes an apparent noise penalty. At the same time, CMS consists of a single range estimate just like in full duplex two-way ranging resulting in the same factor of 2 noise variance benefit compared to TWTT. Ignoring the impact of the transmitter and receiver transfer functions for simplicity, the effective E_s/N_0 for TWTT is

$$\left(\frac{E_s}{N_0}\right)_{TWTT} = \frac{\overline{s_r^2}}{\overline{n^2}} \cdot \alpha m \qquad (7)$$

where α is the number of code copies averaged and m is the code length. The time of arrival is not estimated at node B in CMS, and the signal sent from B to A contains noise from the first leg of the trip. For CMS, then, E_s/N_0 is

$$\frac{E_s}{N_0} = \left(\frac{E_s}{N_0}\right)_{TWTT} \cdot \frac{\alpha}{\overline{n^2} + \alpha} \qquad (8)$$

under the constraint that

$$s_r^2 + \overline{n^2} = 1.$$

The last factor in (8) represents the noise penalty of CMS verses TWTT. This term is unity at infinite SNR because there is no penalty (processing gain provides no benefit without noise). At very low SNR ($\overline{n^2} \approx 1$), the penalty term is approximately ½ if no averaging is used ($\alpha = 1$). The worst case performance degradation is at low SNR, and this factor is cancelled by the factor of 2 difference between the TWTT averaging effect and the CMS single measurement effect. For moderate to large values of α, the penalty term approaches unity (no penalty). CMS with averaging provides better noise performance than TWTT, and it is easy to avoid the sampling penalties common in TWTT.

After a single measurement the variance, σ_r^2, for range binning limited TWTT is given by

$$\sigma_r^2 = \frac{c^2}{12 f_s^2}.$$

Comparing this to the CMS bound, given by

$$\sigma_r^2 = \frac{c^2}{16\pi^2 B^2 \, E_s/N_0}, \qquad (9)$$

we find that CMS has an improved single measurement variance.

$$\frac{\sigma_{r,CMS}^2}{\sigma_{r,TWTT}^2} = \frac{3 f_{sample}^2}{4\pi^2 B^2 E_s/N_0} \qquad (10)$$

Substituting for f_{sample} the factor βB where β represents how much faster the sampling is than the signal bandwidth, we find that if

$$\beta < 2\pi\sqrt{E_s/(3N_0)} \qquad (11)$$

then CMS provides better performance than TWTT. For example, with E_s/N_0 of only 0 dB, β must be 3.6 (where the Nyquist rate is β = 2). At E_s/N_0 of 10 dB, β must be 11.5. This result is directly in line with Figure 2.4 where signals must be highly oversampled to achieve performance approaching the CRB unless CMS is used.

3.1.5 Signal Designs for IEEE 802.15.4 and IEEE 802.11b

Two standards that are relevant to wireless networking are IEEE 802.15.4 for wireless sensor networks and IEEE 802.11b for wireless LAN. The 802.15.4 standard is intended for low data rate (250kbps) wireless sensor networks, and the signal occupies a 2MHz RF bandwidth. The 802.11b standard is widely used for local area networks but at higher data rates (11Mbps) and occupies 11 MHz of RF bandwidth. Table 3.2 summarizes characteristics of these protocols.

The first thing to consider in signal design is the time duration of the signal to ensure unambiguous ranging. In wireless sensor networks, inter-node distances are

Standard	Data rate	Bandwidth	Typical SNR
802.15.4	250 kbps	2 MHz	>6 dB
802.11b	11 Mbps	11 MHz	>10 dB

Table 3.2 IEEE wireless standard summary

Standard	P	t_s	CRB (m_{rms})	β Req. for TWTT
802.11b	8	$727ns$	0.24	58
802.15.4	64	$1\mu s$	0.75	32

Table 3.1 Ranging signal parameters, the resulting Cramér-Rao Bound, and the over sampling rate, β, required in two-way time transfer to achieve the Cramér-Rao Bound

highly variable and outdoor networks up to 100m are reasonably common. In order to accommodate these long links, the signal duration, T_c, must be greater than or equal to $667ns$. Systems with longer link requirements need longer signal durations as discussed in 3.1.3 or need localization algorithms that can intelligently deal with the range ambiguity problem. The chip time in 802.15.4 is $500ns$, and 2 chips are required to exceed the minimum limit. In 802.11b the chip duration is 91 ns, and 8 chips are required to exceed the minimum limit.

Noise performance requirements can be considered now to determine if increasing the duration of the signal is necessary. Although not discussed here, the CRB depends on the modulation scheme because the shape of the spectrum impacts accuracy. The modulation schemes used here are all approximately quadrature phase shift keying (QPSK), and the result in section 2.1 is valid for both 802.15.4 and 802.11b. Because the baseband SNR is large in both cases, the CRB is a good bound on the noise performance of a ranging system. After combining (2.1) and (2.3) and solving for $t_s B$, the required pulse compression factor, P, can be calculated:

$$P = t_s B \geq \frac{c^2}{8\pi^2 B^2 SNR \sigma_{\hat{r}}^2}. \qquad (12)$$

The required values for P to achieve better than a 1m CRB are 1 and 64 for 802.11b and 802.15.4 respectively. In the case of 802.11b, the signal can simply be 8 chips long with no averaging of multiple copies. In 802.15.4, a 2 chips sequence must be averaged 32 times to achieve the required performance. Table 3.1 shows the CRB for these signals showing that short signals are required to achieve reasonable noise limited performance along with the value of β that would be required to achieve the same performance using TWTT.

3.2 Multipath Error Reduction Using An Unbiased Demodulator

Multipath propagation can contribute significant errors to a ranging system operating in a cluttered environment. The density of the clutter required to cause large errors depends on many factors in the signal design and receiver design. In Chapter 2 we discussed how a severe multipath environment is difficult to deal with under any circumstances and how the available mitigation techniques are difficult to implement and are not expected to provide the required accuracy. In this section we discuss a multipath mitigation technique that requires minimal user processing and relies on the inherent properties of the multipath environment and a signal demodulator.

3.1.1 The Demodulator Structure

The signal demodulator used in this system is a simple, digital frequency detector. The standard receiver setup for 802.15.4 is to have a low intermediate frequency (low-IF)

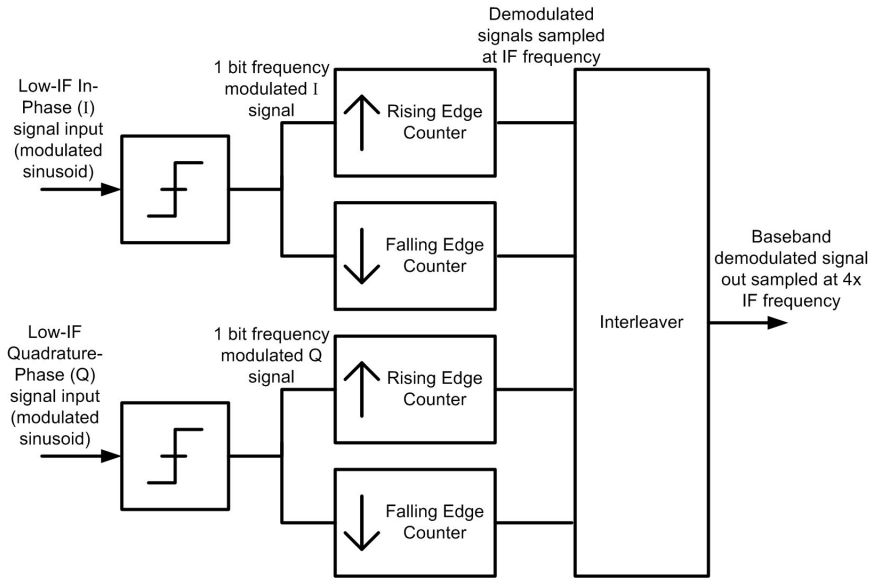


Figure 3.3 Block diagram of low-IF FM demodulator. The slicer block takes a continuous time and amplitude signal and turns it into a continuous time signal with binary levels. The counters measure the time between rising (or falling) edges. Note that the demodulated signals before the interleaver are sampled at ¼ the output rate and each line contains samples offset by ¼ the IF period in time.

receiver with an FM demodulation at the low-IF. Figure 3.3 shows a block diagram of the demodulator used here including all phases. The incoming signal is a modulated sinusoid. It is passed through a slicer (1 bit digitizer), and the period of the resulting square signal (rising edge to rising edge & falling edge to falling edge) is measured using a high speed counter. The count at the end of each period is applied to a lookup table where the count is translated into a demodulation value. This structure is extremely simple, produces a multi-bit frequency estimate, and has reasonably good noise performance. The important things to note here are that the demodulator is only concerned with phase and frequency changes of the signal, and the amplitude is ignored. An example signal input and output are shown in Figure 3.4.

3.1.2 Analysis of Two Path System

The simplest multipath situation is where there is a direct path and a single other path that arrives with some time delay t_d and some carrier phase ϕ relative to the direct path. This situation is actually difficult to understand using hand analysis, but a brief

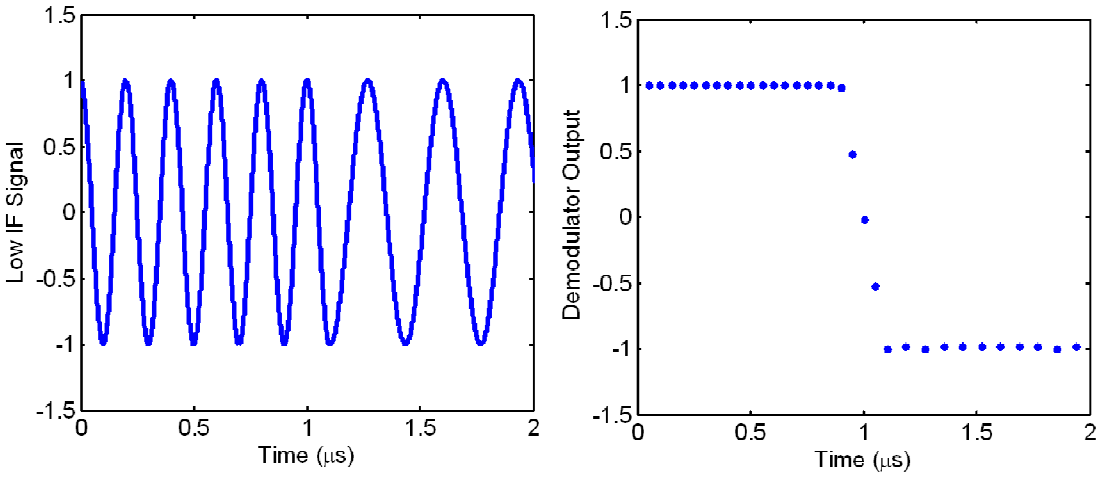


Figure 3.4 Plots showing the analog, modulated low-IF input (I phase only) and the resulting demodulator output generated using both I and Q phases.

analysis that shows why this is the case follows.

For the following calculations, the modulation scheme is frequency shift keying (FSK) with a deviation much smaller than the center frequency. In FSK the modulation signal is square and changes between two frequencies ω_1 and ω_2 and can be represented by the function $\omega(t)$. The direct path signal at the receiver, is

$$s(t) = A\sin\left(\omega(t)\cdot\left(t+\frac{l_d}{c}\right)+\psi\right).$$

The distance between the transmitter and receiver is l_d, the speed of light is c, and the phase contributed by the receiver's phase mismatch with the incoming signal is ψ. The second path is

$$m(t) = B\sin\left(\omega(t)\cdot\left(t+\frac{l_m}{c}\right)+\psi+\phi\right).$$

The total distance traveled by the indirect path is l_m and the additional phase contributed by the reflection is ϕ. Consider that $l_m > l_d$ for all cases. The following simplifications are then helpful:

$$t_d = \frac{l_m - l_d}{c}$$

$$\theta = \psi + \omega(t)\cdot\frac{l_d}{c}$$

We are interested in the difference between these two signals, and the common terms in the arguments can be ignored. To this end, θ can be ignored (set to 0) because this term is common between the two paths. Therefore, changing frequency (through modulation or changing center frequency) will have no impact on our observation. Another way to look at this is that the ϕ term fully captures the phase relationship between these two signals. We can then rewrite $s(t)$ and $m(t)$.

$$s(t) = A\sin(\omega(t)\cdot t)$$

$$m(t) = B\sin(\omega(t)\cdot(t+t_d)+\phi)$$

The total signal at the receiver is

$$r(t) = s(t) + m(t)$$

$$r(t) = A\sin(\omega(t)\cdot t) + B\sin(\omega(t)\cdot(t+t_d)+\phi)$$

This expression is not easily simplified without making simplifying assumptions. Assuming t_d is less than T_{chip}, then there are times when $r(t)$ consists of a sinusoid with a single frequency component. At these times the two path are at the same frequency and are adding together linearly, therefore only a magnitude and phase change is possible. In this portion of the signal, the signal is

$$r(t)|_{\omega_1} = \sqrt{A^2+B^2+2AB\cos(\phi)}\sin\left(\omega_1 t + tan^{-1}\frac{B\,sin(\phi)}{A+B\,cos(\phi)}\right)$$

$$r(t)|_{\omega_2} = \sqrt{A^2+B^2+2AB\cos(\phi-\omega_1 t_d)}\sin\left(\omega_2 t + tan^{-1}\frac{B\,sin(\phi-\omega_1 t_d)}{A+B\,cos(\phi-\omega_1 t_d)}\right)$$

From these expressions, we see that there is a magnitude change between the two signals as well as a change in steady state phase and RF frequency. The most interesting portion of this signal, however, is in the period of time between when these expressions are valid. In this transition region, $r(t)$ is the sum of two sinusoids at different frequencies, and this sum is not necessarily another sinusoid.

$$s(t) = A\sin(\omega_2\cdot t)$$

$$m(t) = B\sin(\omega_1\cdot(t+t_d)+\phi)$$

$$r(t) = A\sin(\omega_2\cdot t) + B\sin(\omega_1\cdot(t+t_d)+\phi)$$

During this time the multipath interference causes magnitude and phase changes that are rapid (yet continuous), and these changes have unknown (non-sinusoidal) characteristics that may impact the demodulator output.

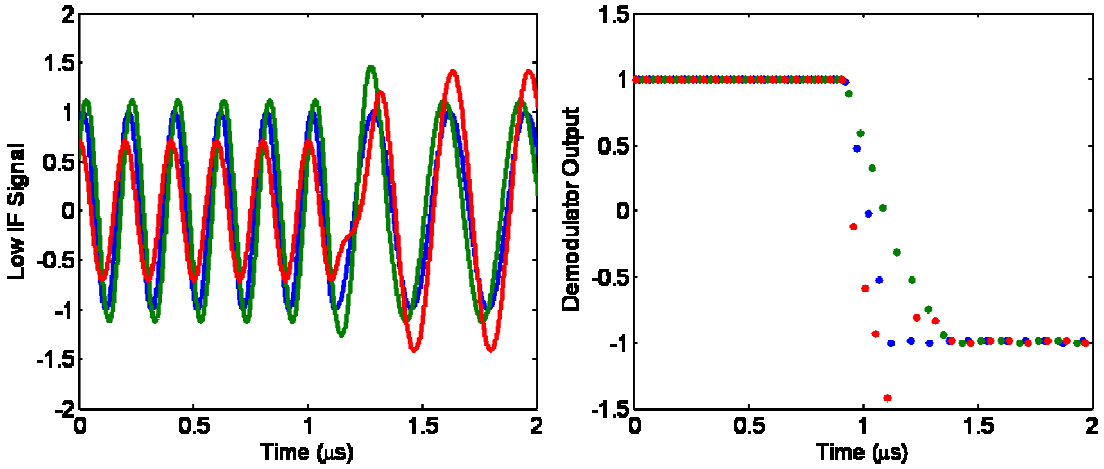


Figure 3.5 Plots showing the impact of a 2-path system on the proposed demodulator. The analog, modulated low-IF inputs (I phases only) are on the left, and the resulting demodulator outputs generated using both I and Q phases are on the right. The blue signal is without multipath. The green signal is with multipath with the relative phase between paths set for maximum error. The red signal is with multipath with the relative phase between paths set for minimum error. The multipath is ½ the amplitude of the direct path and is delayed by 250ns (75m).

Numerical calculations for the combined signals with different phase relationships provide some insight into the effect of the multipath signal. The left half of Figure 3.5 shows the signals before the demodulator. The blue signal is without multipath. The green signal has both paths combined before demodulation, and the multipath signal has half the amplitude, is delayed by 250ns, and has relative phase set for maximum error. This signal, when passed through the demodulator produces the output that is shown in in green on the right half of Figure 3.5. It is clear from looking at the blue and green signals that a positive bias results from this multipath case. Figure 3.5 also shows this situation but with the multipath phase set for minimum error (red), and the unusual behavior is clearly shown. A negative bias results in this case, and this occurs because of the unusual phase behavior around the transition region. This behavior occurs because there are changes to the amplitude, phase and frequency all in a short period of time resulting in a non-intuitive demodulator output.

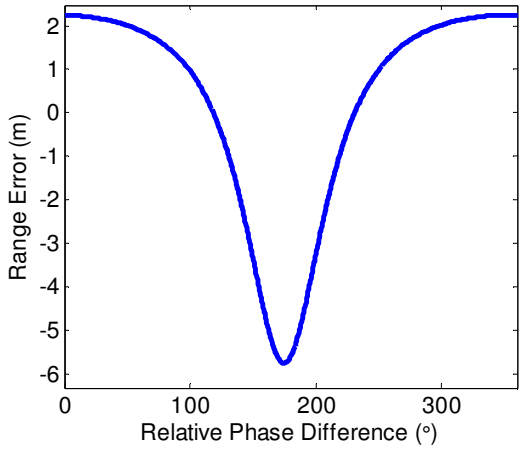


Figure 3.6 Plot showing how varying the phase between the direct path and the second path for a fixed relative amplitude and delay impacts the range estimate. The second path has ½ the amplitude of the direct path and is delayed by 20ns (6m).

The multipath induced bias is a function of the relative multipath amplitude, phase, and delay, and the bias can be positive or negative depending on the relative phase of the paths. This is a very important fact because it is intuitive to believe that only positive biases are possible because the multipath signals always arrive after the direct path. Both positive and negative biases are possible using the proposed demodulator structure because the signal transition areas have very complex phase characteristics due to the multipath channel. Figure 3.6 shows how changing phase impacts the resulting bias estimate for a delay that is much less than the low-IF period for different relative delay values. The amplitude and relative delay also play an important role, and the trends associated with these variables are shown in Figure 3.7. For instances when the multipath amplitude is larger than the direct path, the results are always biased positive. As the relative delay increases, the magnitude of the bias increases up to a point. Eventually the delay is large enough that it can be differentiated from the direct path, and the error decreases. For delays larger than one over the bandwidth, the magnitude of the error is a decreasing function of delay. The trends

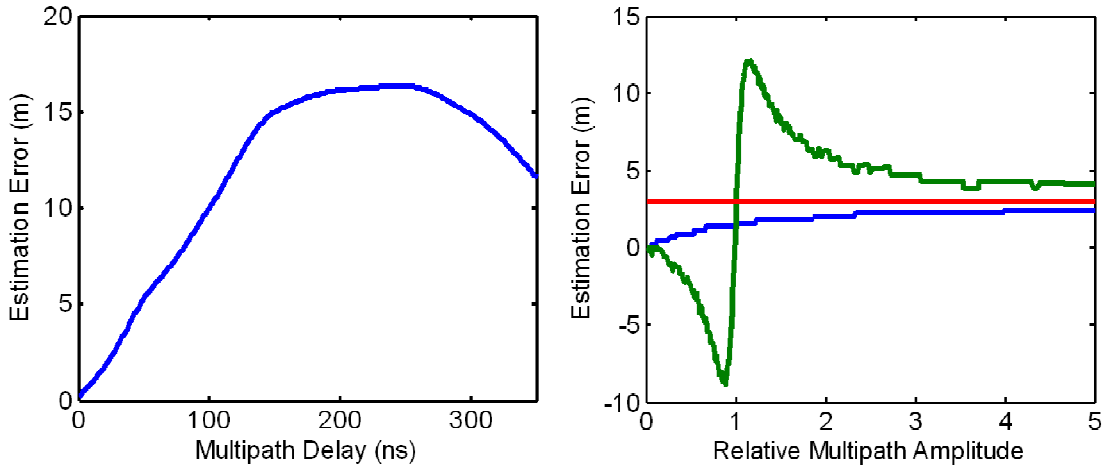


Figure 3.7 Plots showing the impact of varying multipath signal parameters on range estimation error.

Left: Plot showing how range estimation error changes with multipath delay between the direct path and the second path for fixed relative amplitude of ½ and phase set for maximum error. For relatively short delays (less than 120ns), the relationship between delay and error is linear. For larger delays approaching 1/bandwidth, the error levels off, starts reducing, and eventually settles to zero.

Right: Plot showing how range estimation error changes with relative multipath amplitude. The delay is fixed to 10ns (3m) and the delay is shown by the red horizontal line. The blue line shows the error with the multipath in-phase with the direct path (maximum error for very short delay), and the green line shows the multipath anti-phase with the direct path (minimum error for very short delay). For relative amplitudes of greater than 1, the estimation error is always biased positive.

shown in these figures for very short delays (less than about 30ns or 10m) are well described by considering a sort of phasor sum of the signals.

$$Error \approx \frac{A_{MP}\tau_{MP}\cos(\phi)}{A_{MP}\cos(\phi)+A_{DP}}$$

For these short delays and when the signals are in-phase ($\phi = 0$), the linear relationship shown in the left part of Figure 3.7 is clear. For either increasing multipath amplitude, A_{MP}, or increasing multipath delay, τ_{MP}, the error increases. The direct path amplitude is A_{DP}. For larger delays where the delay is similar to the IF period, the relationship is more complicated and cannot be described this easily. The maximum error is no longer due to in-phase signals, and the minimum error is no longer due to anti-phase signals. The general trends, however, remain the same. For some relative

phases, negative biases occur, and as delay increases so does the magnitude of the biases. For relative multipath amplitudes greater than 1, the bias is always positive.

From the trends in Figure 3.6 and Figure 3.7 and in similar situations, it is instructive to consider how to best estimate the true time of flight when presented with a series of measurements taken over the same channel with different phase relationships. To generate these measurements with different phase relationships, the user can take measurements at different carrier frequency, and we discussed how changing carrier frequency changes phase in section 2.4. From looking at Figure 3.6, the mean value may be a good way to approach the unbiased estimate. Other methods may provide better results across a wider set of situations, but the idea that a better estimate can be made by choosing a value closer to the center of the distribution of measurements is instructive. We consider two primary schemes to turn several estimates with different phase relationships into a single, more accurate estimate. The first is the mean method. This method involves taking the mean of several measurements with different phase relationships. The second is a percentile based approach where the 50th percentile represents the median estimate from the estimates

	Mean Error (m)	Error, Percentile (m)								
		10th	20th	24th	25th	26th	30th	40th	50th	75th
RMS Error	5.8	8.3	4.6	4.8	4.3	4.4	5.0	6.6	7.6	9.3
Mean Error	2.7	-6.0	-1.9	0.3	-0.5	-0.8	0.8	3.0	4.3	6.0

Table 3.3 Listing showing root mean square ranging error and mean ranging error under a variety of multipath delays and amplitudes. The delays are 10ns, 20ns, 50ns, 100ns, and 200ns. The relative multipath amplitudes are 0.1, 0.2, 0.5, 0.75 and 0.9. The mean error column represents the mean error across all phase differences for a particular multipath amplitude and delay. The percentile columns represent the value for which X% of the measurements have a value less than given. The rows represent the RMS and mean error across all 25 combinations of delays and amplitudes. The 25th percentile column has the best performance.

with different phases. A summary of the total root-mean-square error and mean error for several different amplitude/delay combinations using either the mean or percentile methods are given in Table 3.3.

From this table we see that the mean is biased positive, but that negative values are still generated. Due to this positive bias, the percentile method provides a better method for estimating the true range estimate, and the 25th percentile in particular is a good estimate of range for the two path case.

3.1.3 Multipath General Case

The general case for a multipath environment is that several paths exist in the channel between the two nodes, and the relationships are not as simple as those discussed for the two path case. Adding additional paths results in even more complicated phase changes, but the general ideas explored in section 3.1.2 remain to be true. Changing the phase relationships between the different paths results in both positive and negative biases, and the general conclusion from Table 3.3 suggests that the best range estimate results from taking the 25th percentile of the time of flight estimates taken at different carrier frequencies. We will use this method for the more complicated environments we see in real environments.

Chapter 4

Prototype Ranging System

The proposed ranging system is a combination of new algorithms that require custom hardware to implement. In order to demonstrate these ideas, a software defined radio platform, dubbed Waldo, was developed. This platform consists of a 2.4 GHz radio, digital to analog interfaces, an FPGA, a microcontroller and the corresponding Verilog and embedded C code required for correct system operation. This chapter will discuss the hardware design along with the general structure and contents of the code. The chapter will conclude with comments on how to improve the design in a future revision.

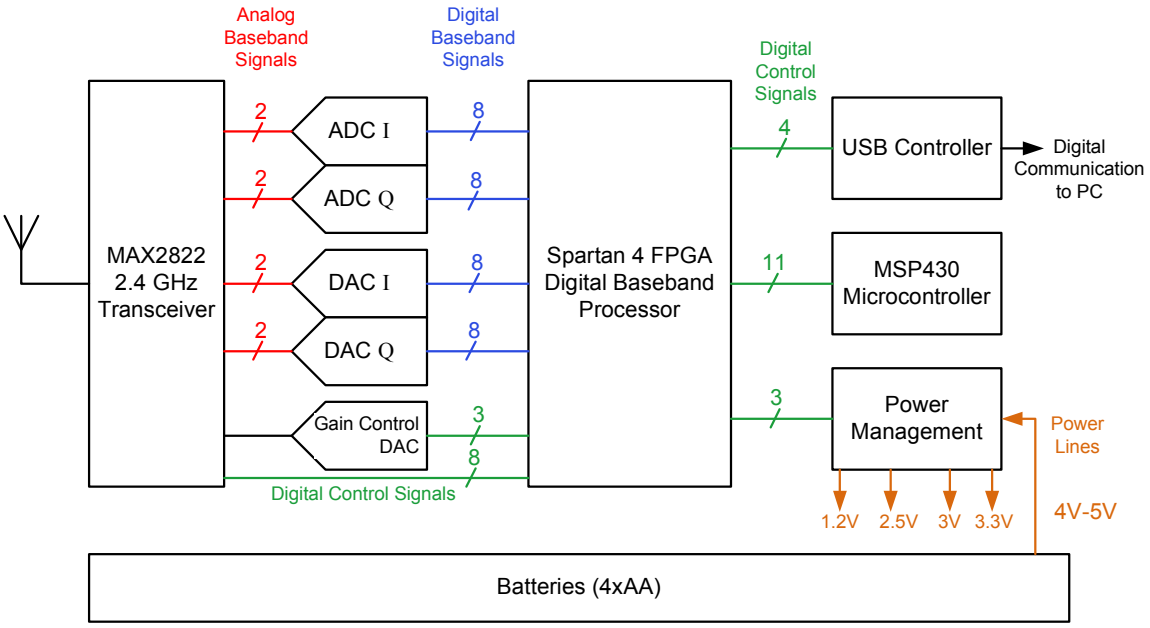


Figure 4.1 Block diagram of the Waldo platform showing the major components. The signal ADCs and DACs are sampled at 25 MS/s and have differential inputs/outputs.

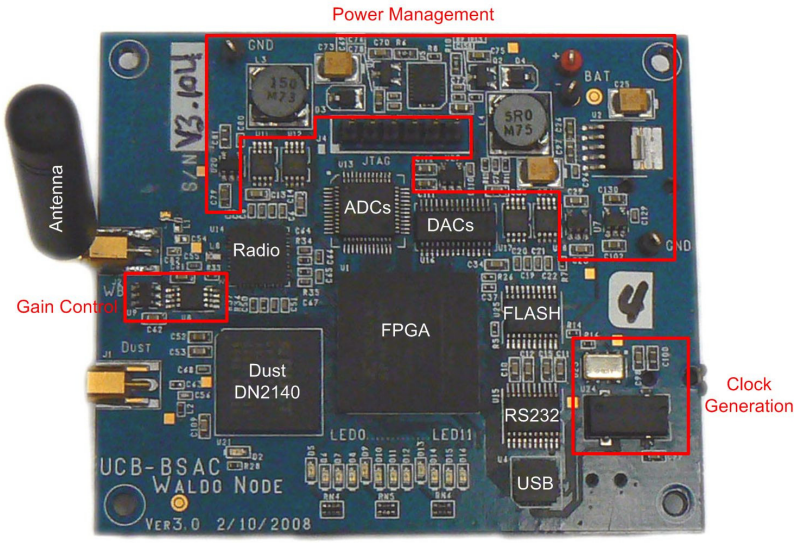


Figure 4.2 Photograph of the Waldo board with the components labeled. The board is 7.5cm x 6cm, and all of the major components are on the top side. The FLASH chip is used to program the FPGA. The Dust DN2140 and RS232 components are unused in the current implementation.

4.1 WALDO HARDWARE OVERVIEW

The Waldo hardware platform is shown in block diagram form in Figure 4.1. Each component was chosen to maximize the flexibility of the platform while providing for a low cost, compact platform that can be deployed in networks for several hours at a time. The Waldo platform was not designed to be a low power platform like typical node platforms, and it was not intended to be fielded for months at a time without battery changes. The power consumption of the platform, however, was taken into consideration during the design process to ensure that Waldo could run for hours without a battery change. Similar software defined radio platforms consume amps of current, require a PC at each node to operate (GNU radio, [34]), and are contained in large boxes. Waldo was specifically designed to be highly portable and self contained. A photograph of the Waldo board is shown in Figure 4.2.

4.1.1 RF Transceiver

A flexible 2.4 GHz radio, the Maxim MAX2822, was chosen for Waldo to ensure that a variety of physical layer implementations are possible. The radio standards of interest for this work are IEEE 802.15.4 and 802.11b. Therefore, a radio capable of meeting these standards was required. Most radio platforms do not allow for a variety of physical layer changes because they are so called "data in, data out" transceivers. That is, you provide a packet of digital information to the radio, and it transmits it using a predefined modulation scheme and method. When an appropriate RF signal is received, it demodulates it using a specific method and produces a packet of digital bits to be passed to the rest of the system. The latencies associated with this process and the lack of direct baseband information make ranging difficult using a transceiver of this type. The radio chosen has analog baseband inputs and outputs for both the in-phase (I) and quadrature phase (Q) signals, and this enables us to implement virtually any modulation scheme. The transceiver is half-duplex (it can either transmit or receive but not both at the same time), and both the transmitter and receiver share a frequency synthesizer. The synthesizer is implemented with an RF local oscillator and an integer-N phase locked loop that has 1 MHz channel spacing.

The receiver is designed as a direct conversion receiver for 802.11b, and each baseband signal has a 7 MHz bandwidth. This configuration provides a total RF bandwidth of 14 MHz. The 802.11b standard uses about 11 MHz of bandwidth, and the 802.15.4 standard uses about 2 MHz of bandwidth. We use this receiver at a low-IF of 4 MHz, and the signal bandwidth is roughly from 3 MHz to 5 MHz. The receiver has variable gain and a linear receiver chain, but there is no analog image rejection or

channel filtering (beyond the 7 MHz low pass filter) in this configuration. There is an analog input that controls the receiver gain.

The transmitter is designed as a direct conversion transmitter for 802.11b, but there is a baseband filter at 10 MHz to limit the transmit bandwidth. We use this as a low-IF transmitter at a 4 MHz IF. Image rejection is implemented by the transmit up-conversion mixers. The nominal transmitter power is +14dBm (25 mW).

The transceiver has a high degree of digital reconfigurability. The transmit/receive state is controlled by direct digital pins. Many other features (including synthesizer frequency) are controlled through a 3-wire SPI interface.

4.1.2 Analog to Digital Interfaces

The transmitted and received baseband (low-IF) signals are analog, and they are generated or digitized using digital to analog converters or analog to digital converters respectively. To ensure that the full bandwidths could be achieved, the parts were chosen to have relatively high sample rates. The resolution of the parts, however, is just 8 bits. Many RF systems use higher resolution converters, but the available dynamic range in these parts is sufficient from a noise standpoint. The primary issue with these low resolution parts is the resulting transmitter spectrum has a great deal of out of band spectral power and relatively poor image rejection. For a prototype system, these characteristics are tolerable, but out of band RF performance could be improved significantly with a few extra bits.

The digital to analog converter is a MAX5189 with a maximum update rate of 40 MSps. The analog to digital converter is an Analog Devices AD9288 capable of sampling at up to 100 MSps. Both parts are operated at 25 MSps in the system. The maximum

signal bandwidth is about 10 MHz, so this sampling rate is capable of operation without aliasing.

4.1.3 Field Programmable Gate Array

The output of the analog digital converter and the input of the digital to analog converter are connected to an FPGA, and this part is used to implement the remainder of the physical layer. The chosen part is a Xilinx Spartan 4 XC3S1000 FPGA, and it contains 1 million equivalent gates. The main clock is 100 MHz, and a 32 kHz clock is available for lower power operation.

4.1.4 Microcontroller

The Waldo platform was not originally designed to have a microcontroller because the FPGA firmware would contain all of the required logic and control. FPGA code is much more time consuming to produce and debug than embedded C code. It became clear that non-time-critical control and processing steps could be carried out equally well in a

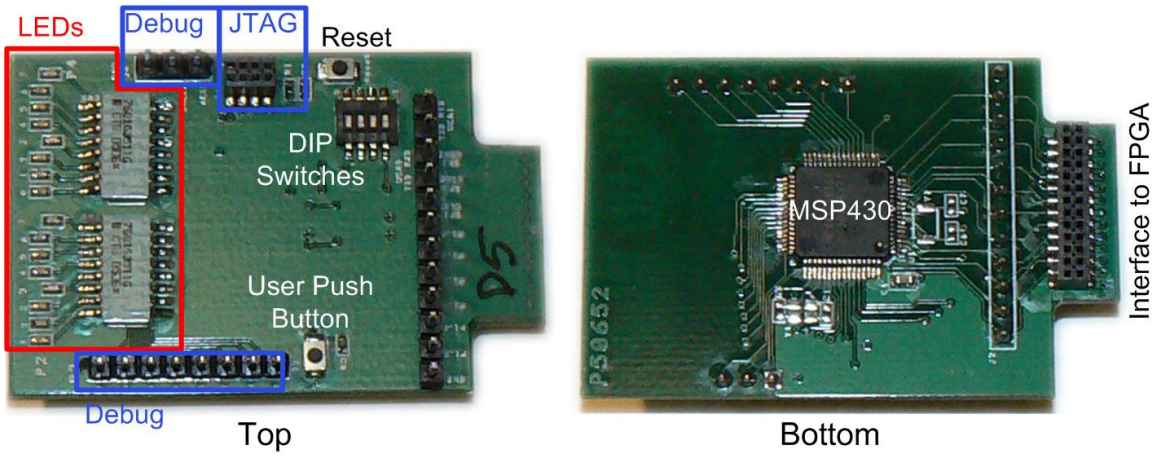


Figure 4.3 Photos of Waldo daughter board showing top and bottom views. The daughter board mounts to the back of the main Waldo board and adds a microcontroller, DIP switches and additional LEDs and pins for debugging.

microcontroller with a far shorter development time. A daughter board was added to Waldo that contained a microcontroller, and a total of 11 digital pins connect the two boards together. The daughterboard includes a few nice additions to the system other than the microcontroller. It includes 16 additional LEDs for debugging, a user push button switch, and a 4 bit DIP switch. A photograph of the daughterboard is shown in Figure 4.3.

An MSP430 microcontroller was chosen for the microcontroller primarily because others in the research group have had experience developing software for this family of parts. The MSP430 was chosen to have a large amount of RAM and FLASH to ensure that development was not slowed by microcontroller resources. This part is a 16 bit microcontroller capable of running at 16 MHz with 92kB of FLASH memory and 8kB of RAM. It includes many useful peripherals for embedded systems, and is an extremely low power device given its processing capability and many sleep modes. The low power capability is important in a general embedded system, but it is not critical here because of the high power consumption of the Waldo board itself.

4.1.5 Power Management

The Waldo platform has several different power domains, and it is possible to turn off power to most devices. The power management structure consists of a switching buck converter that takes in the battery voltage and generates 3.3V and 2.5V and 1.2V. The 3.3V output is then linearly regulated to 3V to run the radio and other analog components. Waldo is powered by 4 AA batteries (typically nickel-metal-hydride batteries) for an input voltage of about 5V. The maximum input voltage is 6.5V. When the battery voltage drops to 3.4V, the system stops functioning, but the batteries are highly discharged by this point (Cell voltage of 0.85V). The efficiency of switching

Waldo Mode	Power
Radio Off	175 mW
Receive Mode	1145 mW
Transmit Mode	1440 mW

Table 4.1 Waldo power consumption in three modes of operation including the microprocessor.

Condition	Lifetime
0% Duty Cycle (Idle)	14 Hours
1% Duty Cycle	13 Hours
5% Duty Cycle	11 Hours
10% Duty Cycle	9 Hours
100% Duty Cycle	2 Hours

Table 4.2 Lifetime of a Waldo platform for varying radio duty cycle when powered by 4 NiMH AA batteries (9000J).

converter is between 80% and 90% over the range of common usage thereby ensuring efficient use of the battery energy. Although many features were included to reduce power consumption, this platform is not a low power platform. Software defined radio systems like Waldo can be used for high flexibility at the cost of high power consumption. As a result, communications and ranging for embedded systems can be developed using software defined platforms such as this, but the end deployment should be done using an application specific integrated circuit approach rather than a software defined approach. The power characteristics of the Waldo platform are shown in Table 4.2 and Table 4.1.

4.2 WALDO SOFTWARE OVERVIEW

All of the Waldo nodes not only have identical hardware, but they have identical software. Each node is set as a base node (one that originates the ranging signal) or as a repeater node (the node that receives and then replies with the ranging signal) through the use of commands sent over the wireless link. The software system implemented for Waldo was written for three different platforms using three different languages. The FPGA software is written in Verilog using Xilinx extensions (the firmware), the microcontroller code is written in MSP430 embedded C (the software), and the highest level of network control and localization is implemented in Matlab on a PC. Figure 4.4

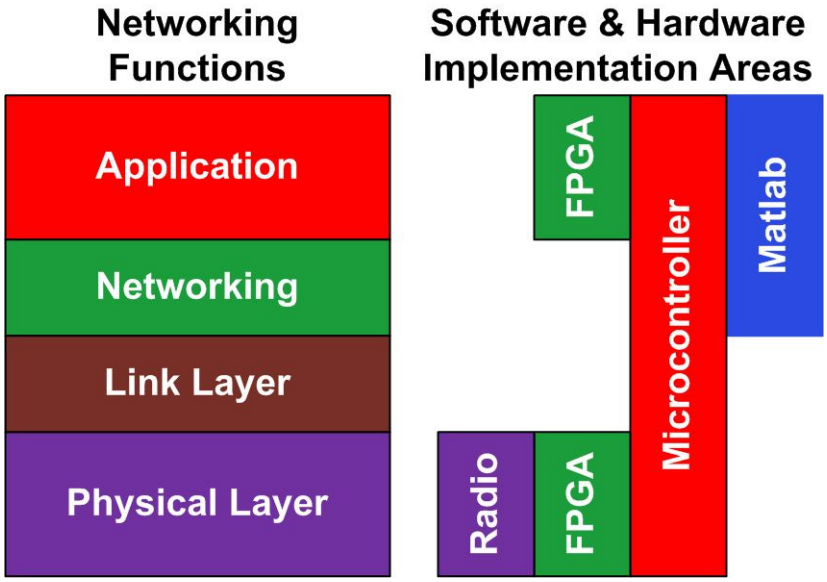


Figure 4.4 Diagram showing how the software and networking functions (left) are divided across the hardware and software resources available to the system (right). The FPGA implements application level functions for ranging but not communication. The microcontroller implements physical layer functions for communication but not ranging.

shows generally how the overall system software is divided up across these three platforms. The FPGA handles most of the physical layer, but part of the data communication physical layer is in the microcontroller. The microcontroller also handles medium access and application issues. There is no true networking layer (or the other missing layers) in the Waldo system because the overall network structure is extremely simple and centrally controlled. This section contains an overview of the various software components of the system starting with the firmware on the FPGA, followed by the embedded software and then the Matlab code. The FPGA main clock is 100 MHz, but the system core runs at 25 MHz. The microcontroller runs at 16 MHz.

4.2.1 Low-IF Transmit Chain

The radio takes I and Q baseband signals and mixes them up to RF, and these signals are generated using the low-IF transmit chain. Before reaching the radio, the signals generated by this block are converted to analog using a pair of digital to analog

converters. The generated signals are at a 4 MHz low-IF and are nominally 90º out of phase. This section consists of two lookup tables and some logic that translates the input digital word into the correct signals for these tables. The input digital word is 5 bits and is sampled at 25 MHz. A 0 corresponds to 2 MHz output I and Q signals with a programmable phase offset, and a 31 corresponds to a 6 MHz output. By default the phase difference is set to 90º. One of the two tables is a sine table with 1024 entries, and the second table converts the incoming digital word to the distance between entries in the sine table for the desired output frequency. For example, at a 25 MHz update rate, a change of 1 entry per cycle would result in an output frequency of 24.4 kHz. To output 4 MHz, ideally a step of 163.8 entries per cycle is used. Fractional cycles are not handled in this system for simplicity, and the modulation step size is an average of 129 kHz per LSB of the digital input word. This is a large step size, and it would be better if it were reduced significantly. There is no reason to be able to generate such a wide range of frequencies, so this could be improved. It was designed to correspond with the demodulator to be described in the next section for simplicity, and the resolution of the demodulator is fundamentally limited by the maximum system clock frequency.

4.2.2 Low-IF Receive Chain

The majority of the receive chain is implemented in the FPGA firmware. In this part of the firmware, the received signal passes through a band-pass filter, a 4x interpolation filter, and an FM demodulator. There is no image rejection performed in the digital domain even through the information to do so is available. This choice was made for simplicity at the expense of noise and interference performance.

The incoming I and Q signals are at a low-IF of 4 MHz, and the signals are low-pass filtered by the radio hardware to have a bandwidth of 10 MHz. These signals are

sampled by a 25 MSps, 8 bit analog to digital converter. The incoming 8 bits per sample are not all used for communication or ranging, but the 5 most significant bits are passed through a 25 MSps IIR (Chebyshev type I) band pass filter to reduce the received signal bandwidth to 2 MHz. This filter is implemented with approximately 8 bit fixed point coefficients, and it is a custom implementation. Xilinx provides an automatically generated filter core, but it uses a block memory and hardware multipliers. The custom implementation has some coefficients that are more precise and some that are less so, but the data path width is 10 bits. This combination of variable width coefficients along with a 10 bit data path provides for a compact implementation without the use of hardware multipliers or memories. The number of input bits, coefficient bits, and data path width were chosen using Matlab to consider the tradeoffs. This was one of the first blocks implemented, and it was over-designed in terms of resource savings.

The filtered data then moves through a 4x interpolation filter. This filter increases the sample rate by 4x to improve the time resolution of the signal zero crossings. In the demodulator section, the importance of this improvement in resolution will be clear. The interpolation filter was implemented using a Xilinx core for simplicity. Both the I and Q channels are filtered this way.

The FM demodulator rounds the demodulator output to 1 bit, and measures the time from rising edge to rising edge and falling edge to falling edge. The 100 MHz clock has 28.6 clock ticks during a logic 0 (3.5 MHz into the demodulator) and 22.2 clock ticks during a logic 1 (4.5 MHz). This difference corresponds to 2.7 bits of information regarding the received signal frequency. Without the interpolation filter, only 0.7 bits of information would be available. This difference is significant in that this extra information is very useful for ranging. The demodulated signal using this technique on a

single sine wave has a sample rate of twice the IF frequency because the period is measured rising edge to rising edge and falling edge to falling edge. The demodulator used here, however, uses both the I & Q channels resulting in a sample rate of 4x the IF frequency or 16 MSps in this case. For practical reasons, the output is resampled to 25 MSps to be better synchronized with the system clock. The signal interpolation is only to provide more bits of information per sample rather than to increase the resulting demodulated signal sample rate. The demodulated signal sample rate is set only by the IF frequency. This FM demodulator architecture is standard in custom ICs, and the clock that measures the length of the period is often derived from the local oscillator and is several hundred megahertz. This is easily done on-chip and would result in a greater number of bits per demodulator output sample resulting in higher available baseband (demodulated) SNR. The maximum SNR in this implementation is about 17.8 dB, but operating at 300 MHz (a standard choice as it is 2.4 GHz ÷ 8) would result in a maximum SNR of 28 dB. This 10 dB difference can make a big impact in performance. If the IF period was measured only with the 25 MHz clock, the maximum SNR is only 1 dB resulting in poor performance. The 4x interpolation of the IF signal is critical to enabling reasonable baseband SNR.

The multi-bit frequency estimate is applied to a lookup table where a 2 MHz signal corresponds to 0 and a 6 MHz signal corresponds to 31, and the frequency step size is the same as in the FM modulation block described earlier. In a ranging measurement all 5 of these bits are retained, but this signal is resolved to a single bit using hysteresis for data communication. The resulting 5 bit signal is used in the ranging digital baseband, and the single bit version is used in the communication subsystem.

4.2.3 Multiple Address Accumulator

The multiple address accumulator block implements the core of the code modulus synchronization algorithm used in this ranging system. This block averages the incoming demodulated signal in the same way that an oscilloscope does: each sample point is averaged with sample points from the same relative time in the period of the signal. This method of averaging has the same effect as other methods of processing gain, and the SNR increases linearly with each averaging operation.

The accumulator block is implemented using a dual ported block RAM on the FPGA, and there is some logic associated with the block as well. The incoming signal is added to the current value for that sample in the RAM, and the result is stored back into the RAM. The number of samples per period of the signal is known, and the controlling logic is setup to ensure that exactly a single cycle of the incoming signal is added together several times. This operation is done after demodulation, and the input signal is a band-limited pulse sequence. The current implementation records data at 100 MSps although the input data is sampled at 25 MSps. The incoming signal is 2 μs long

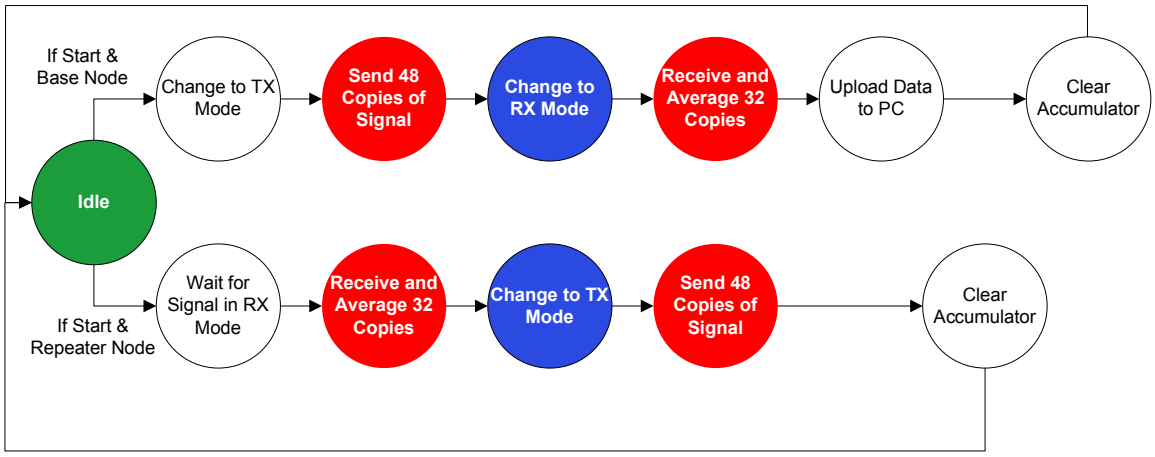


Figure 4.5 Flow diagram showing high level states of the firmware ranging state machine. The red and blue blocks are the blocks where timing is critical and the state machine implements the code modulus synchronization state machine.

resulting in a 200 sample record of the input signal, but only 50 of those samples contain new information.

4.2.4 Digital Baseband Finite State Machine

The ranging operations are initiated by the software on the microcontroller, but the actual ranging operations are controlled by a finite state machine implemented on the FPGA. Both the base node and the repeater node use the same state machine, but the state machine is partitioned into base and repeater sections. The total state machine consists of only 24 states approximately half of which are for the base node and the other half are for the repeater node. There are other state machines in the system that are started by the main state machine, and these were largely (if informally) discussed in Sections 4.2.1 through 4.2.3. The timing of the initiation of multiple ranging operations is controlled through software because the initiation time is not critical on the nanosecond level.

A high level summary of the state machine is shown in Figure 4.5. The state machine waits in an idle mode for a ranging command from the microcontroller. In the idle mode, control over the transceiver state is given to the microcontroller. When a ranging command is received from the microcontroller, it also contains information regarding base or repeater status. The state machine takes control of the transceiver state, and the base node switches to transmit mode, and the repeater is in receive mode. The base node transmits 48 copies of the code, a 2 chip code of 0b10 in this case, and this code is synchronized a local reference clock to ensure code modulus synchronization functionality. The repeater node detects the incoming signal and receives 32 copies of the code in a manner synchronized with a local reference to ensure code modulus synchronization (refer to section 3.15 for choice of 32 copies of a

2 chip code). These 32 copies are accumulated together to effectively average the code 32 times. The transceivers switch states, and the repeater sends the code back 48 times, and the base node receives 32 copies. The choice to send 48 copies comes from a legacy implementation where an 8 chip code was used, and the total length of the transmitted signal was retained for convenience. Both nodes are mindful of the local time reference thus ensuring code modulus synchronization. The repeater then clears the accumulator memory and returns to idle mode. The base node dumps the memory contents to the microcontroller, and then the base node clears the memory and returns to the idle mode.

4.2.5 Wireless Communication Subsystem (Physical and Link Layers)

The wireless communication subsystem is required to send commands and data to and from the base station, and it uses the same low-IF transmit and receive firmware, sections of the microcontroller hardware, and embedded software. The low-IF segments have been described above, and the two additional parts will be described here along with the packet structure.

Timing recovery and bit detection are implemented using a UART on the MSP430 microcontroller (universal asynchronous receiver/transmitter). The low-IF receive chain provides a 1 bit signal out of the demodulator (along with a multi-bit signal) that contains the noisy communication data or ranging signal. This signal is also asynchronous to the system clock, and timing recovery must be performed to recover the incoming data bits. To ease the implementation of timing recovery, the transmitted data packet is constructed and sent using a UART. As a result, each 8 bit byte has a start bit of 0 and a stop bit of 1 appended to ease timing recovery. The receive UART available on the MSP430 microcontroller is superior to the one we implemented in the

FPGA in that it over samples the signal and uses the sum of these samples to make decisions. The Verilog UART used takes a single sample from the center of the bit and makes a decision. This difference results in a few dB improvement in receiver sensitivity. The UART transmits and receives at 1Mbps resulting in a raw data throughput of 800kbps with an additional 200kbps of start and stop bits. This is clearly not an efficient scheme, but simplicity is essential to successful system implementation.

A custom packet structure was developed for this system due to the unique constraints of using a wireless UART and the potential need to send large amounts of data through the network. The packet structure is shown in Figure 4.6.

The preamble is constructed from twelve 0xFF bytes and is much longer than required in the current implementation. This length was chosen to ensure that the preamble could be detected at maximum gain, the automatic gain control loop could settle, and enough preamble would remain for resynchronization if required. It takes up to 3 Bytes to ensure that the signal observed is the preamble. The 1st byte arrives immediately after a byte of noise, and it is not possible to ensure that the start bit is detected as a start bit instead of a data bit. That is, the noise could be mistaken to contain a start bit. If this occurs the UART will see an idle line signal (all 1's) until the start bit of the 2nd byte. It was decided arbitrarily that two consecutive 0xFF bytes were needed to ensure that the observed signal is a preamble and not noise. The automatic

Figure 4.6 The Waldo packet structure. The top line is the field name, and the 1st part of the second line is the length of the field. The 2nd part of the second line is either the value of those bytes or the coding used on the bytes. ECC denotes the use of an hamming (8,4) error correcting code, Raw denotes uncoded data. The final checksum is calculated over just the data bytes.

gain control loop was then intended to run to set the gain appropriately, and this loop sets the gain in $60\mu s$. It was determined that using gain control did not impact packet error rate, and this step was not included in the receiver stack for simplicity. Following the preamble, a single 0x55 byte is included as the start symbol. This byte could be chosen to be any value, other than 0xFF. Up to this point, no bit errors have been considered acceptable, and any single bit error results in a reset of the receive packet state machine. Only a single 0xFF byte is required before the start symbol for a packet to be received, and this results in a requirement that the exact sequence 0xFF55 (excluding UART added bits) be received to start a packet. The probability that random noise will generate the appropriate start sequence is less than 10^{-6} which is acceptable given that additional packet validity checking is performed after this start sequence.

The next three bytes are the packet source, destination and length fields (one byte each). Each byte is (8,4) hamming encoded to decrease the probability of a false packet being received. The notation (8,4) means that 8 bits are sent to encode 4 data bits. The (8,4) hamming encoding provides single bit error correction while still providing detection of 2 bit errors. The hamming encoding is done via a software lookup table on the microcontroller, and the decoding is done in software. Each byte can only specify $2^4 = 16$ values, and this limits the size of the network to 16 devices. This limitation is not an issue for networks constructed at this point, but it is clear that 16 devices is too limiting for a broader application of the technology. The length field is also limited to only 16 values, but this is solved by a translation of length field to actual packet length as follows:

a) A length field of 0 is reserved for packet acknowledgements and specifies an acknowledgement length of 4B (18B including overhead).

b) A length field of $x \neq 0$ specifies a packet of length $16x$ bytes.

The maximum packet length is 256 bytes, and this is a reasonable length given the amount of data that might need to be moved across the network. Two hamming encoded bytes could be used for the length field but it was convenient in software to have the entire header contained in just 16B, and this coarse granularity in packet length is acceptable for this test application. If the hamming encoding is not capable of correcting one of the incoming bytes (because there is more than 1 bit error), the incoming packet is considered to be faulty. The entire packet is rejected if the packet is deemed faulty at any time, and the receiver returns to searching for a new incoming packet. Including the preamble, start symbol, source, destination and length fields, there are a total of 16 bytes of header in the packet.

The next N bytes are data or acknowledgement information, and these bytes are uncoded. Following the data bytes, there is a single checksum byte that is the bitwise XOR of all of the N data bytes. Because of this extra byte, the actual length of a non-acknowledgement pack is $16 + 16x + 1$. The checksum provides a small amount of data integrity checking, and it is primarily intended to reject packets that have a partial collision late in the packet.

The communications software subsystem uses interrupts to ensure the system can continue performing other tasks while transmitting or receiving packets. In the current system implementation, this results in the system having a great deal of spare cycle time. On the transmit side, a packet is constructed in a pre-allocated array, and a transmit packet flag is set. The microcontroller ships a byte out of the UART each time the UART buffer is available. After the packet is sent, the transceiver changes to receive mode and waits for an acknowledgement. If the acknowledgement is received, the

transmit success flag is set, and the system moves on to a waiting task if one exists. On the receive side, incoming bytes are placed into a pre-allocated receive array whenever a new byte is available, and the new received data flag is set. When this flag is high, the software checks the data to see if it is consistent with an incoming packet, sets a task in the next task list, or passes bytes off of the microcontroller to either the FPGA or an attached PC.

4.2.6 Node Control and Networking

The communications subsystem consists of physical, link, and application layers, and the rest of the networking stack is not included for simplicity. This limits the scope of the resulting network, but interesting demonstrations are still possible. As the software is currently implemented, one node is connected to a laptop as the network base station. This node must be involved in all ranging operations and is the source of all commands in the network. The node addresses are set using the 4 bits available on the DIP switch on the microcontroller daughterboard, and these addresses are included in the Matlab control code at run time eliminating the need for network joining. This section describes what amounts to the application layer of this system, and it is a combination of embedded code and Matlab code.

To perform a ranging operation, the PC sends a ranging command to the tethered node. The tethered node forwards this command using a wireless packet to a partner node. The embedded software on the two nodes attempts ranging measurements on each of the 16 pre-specified carrier frequencies. After each measurement, the accumulator data from the base node is dumped from the FPGA and reported back to the PC where it is stored for later analysis. After the 16 ranging operations are complete, the system returns to an idle state.

The command sent from the PC to the tethered node is a 2 byte ranging command sent via USB of the form 0x080P where P is the address of the partner for the ranging operation. The tethered node then sends a wireless packet to node P that is a command to participate in a ranging operation. P replies with an acknowledgement, and ranging operations are attempted on all 16 channels. The timing of these operations must be somewhat carefully controlled, and embedded software is used to ensure that 16 measurements are taken. In the current code, 16 received signal strength measurements are also included, but the results are not reported to the PC due to an unknown bug in the system firmware. These measurements would nominally be performed using the automatic gain control loop. Each measurement is attempted without verification of success. That is, a measurement is tried on each channel, and results are reported to the PC even if the ranging measurement fails. Failures can occur if the channel was occupied and ranging resulted in a collision or if ranging failed due to insufficient link margin. The existence of a failure case is determined after the fact in the analysis software. The RSS hardware functioned in an earlier revision of the firmware that had other bugs, and some data presented in Chapter 5 will include this RSS functionality.

4.2.7 Data Processing and Analysis

All of the data processing and analysis is performed in Matlab. The data processing is conceptually simple and could be performed in the embedded microprocessor, but it was never transferred to the microcontroller because all of the raw data is recorded on a PC to ease debugging and testing of new back end algorithms. The data processing includes determining if the ranging measurement was valid, performing a time of flight estimate based on each valid measurement, and producing the overall time of flight

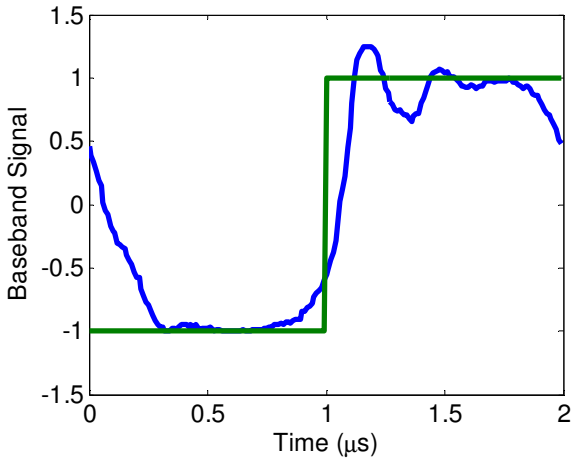


Figure 4.7 Plot of raw received ranging signal after accumulation (averaging) of 32 signal copies (blue) and the template signal that will be correlated against the received signal.

range estimate based on the available data. In localization experiments, a simple localization algorithm is included to convert the multiple ranges into a location estimate.

The data returned to the PC is cross-correlated with a template to determine if it is a valid range measurement and as the first step in range estimation. The result of the cross-correlation is a signal with a relatively sharp peak, and the peak location is a good estimate of the time of flight. The signal that is transmitted in this version of the system is a simple 2 chip sequence of 0b01, and the returned data is correlated against the ideal, square ranging signal 0b01. Using a band-limited template (or even smoothed measured data as a template) rather than the ideal signal does not provide an improvement in ranging accuracy because the band-limited versions contain no additional information regarding the shift of the received signal. The ideal signal is much simpler to process in dedicated hardware, and the analysis was developed with the hardware implementation in mind. Figure 4.7 shows sample data and the template. The correlation that is used is a circular correlation rather than a simple linear correlation. The ranging signal used is periodic and will be circularly shifted due to the

time of flight. We are interested in detecting this circular shift, and a circular correlation is the best detector for this case. Figure 4.8 shows the result of a circular correlation between the data and mask shown in Figure 4.7.

At the end of section 2.3, we briefly discussed how the estimation of arrival time is more complicated than a simple zero crossing detection. The zero crossing detector is not the optimal way to estimate the time of the event because information about the edge arrival time is contained in the entire edge not just the zero crossing. In order to capture as much information is possible, the correlator receiver has been shown to be optimal in white noise (and close to optimal in multipath environments) in the case of range estimation [12]. Circular correlation is used in the analysis of periodic signals whereas linear correlation is used in the analysis of aperiodic signals. A circular convolution essentially slides a periodic template along one period of the received signal, and records the sum of the product of these two waveforms at each offset [35].

This circular correlation can be implemented directly or can be implemented with an FFT. The FFT implementation is useful in that it defines the circular correlation operation in an exact, succinct way. The received ranging signal is given by $s(n)$, the

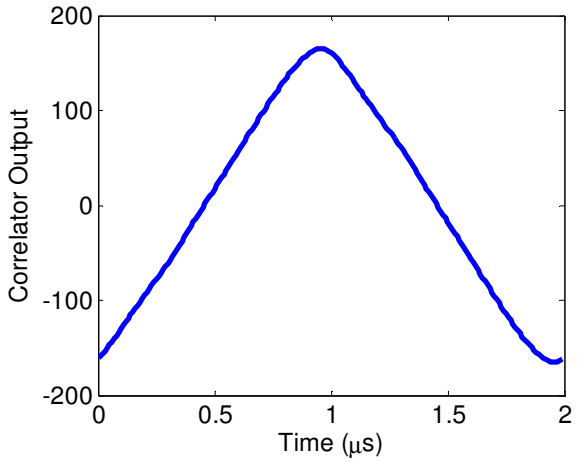


Figure 4.8 The circular correlation between the two signals shown in Figure 4.7.

template signal is given by $m(n)$. The corresponding signals in the Fourier domain are $S(n)$ and $M(n)$. The correlation signal, $r(n)$ in the sampled time domain and $R(n)$ in the Fourier domain, can be found through the multiplication of the Fourier domain versions of the signal.

$$R(n) = S(n) \cdot M(n)$$

This method is computationally efficient in Matlab, but does not lend itself to an embedded implementation because of the large number of complex multiplications and additions. A direct implementation has a greater number of operations, but each operation is extremely simple. First we define the operation $\dot{+}$ to be a circular addition.

$$n \dot{+} k = \begin{cases} n + k & \text{if} \quad n + k \leq N - 1 \\ n + k - N & \text{if} \quad n + k > N - 1 \end{cases}$$

Then, using the circular addition, we can define the circular correlation operation.

$$r(n) = \sum_{k=0}^{N-1} \begin{cases} s(n \dot{+} k) & \text{if} \quad m(k) = 1 \\ -s(n \dot{+} k) & \text{if} \quad m(k) = 0 \end{cases}$$

This operation is extremely simple in that it only involves additions and subtractions, but the number of operations is N^2. This direct calculation need not be done for all N (if

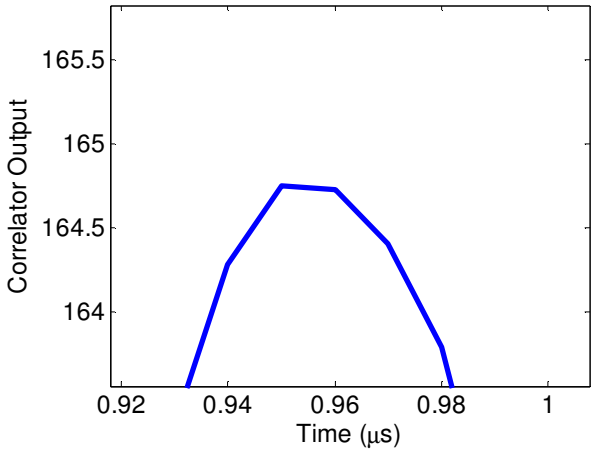


Figure 4.9 Enlarged view of the peak of the correlation result from Figure 4.8. The peak area is flat and estimating the peak location is difficult without considering the more highly sloped regions adjacent to the peak.

the general location of the peak of found using a simple method), and an analysis of the computation complexity of the FFT version on an embedded microcontroller suggests that the direct method will be faster for the data set sizes of interest.

After the correlation has been computed, the peak location must be estimated. This peak location can be determined by simply finding the highest value, but the sample rate of this result is quite low, and the resulting measurement will be limited by range binning. In addition, the peak of the correlation is rounded because of the received signal is band limited, and detection of the actual peak introduces significant errors due to this flatness. Figure 4.9 shows a zoomed in view of the peak of the correlation function, and the flat peak is clearly visible. This figure also shows that the magnitude of the slope increases in either direction from the peak approximately symmetrically, and it may be possible to estimate the peak location using the sloped section of the correlation function. GPS receivers often estimate the peak location using the sloped section of the correlation function, and this method is referred to as the dual correlator peak estimation method [36]. This method attempts to find the value n that solves the following minimization problem where Δ is a fixed number of samples:

$$\min\{|r(n) - r(n + \Delta)|\}.$$

One obvious solution is to try each value of n, and choose the n for which the minimum is achieved. The resulting peak location estimate is $n + \frac{\Delta}{2}$. This method is still limited by range binning but it now avoids the flat portion of the peak as long as Δ is chosen to be large enough.

The peak estimation implemented in this system uses a dual correlator estimate augmented by a simple interpolation scheme to reduce the impact of range binning. The dual correlator estimate is used to find a set of points on either side of the peak that are

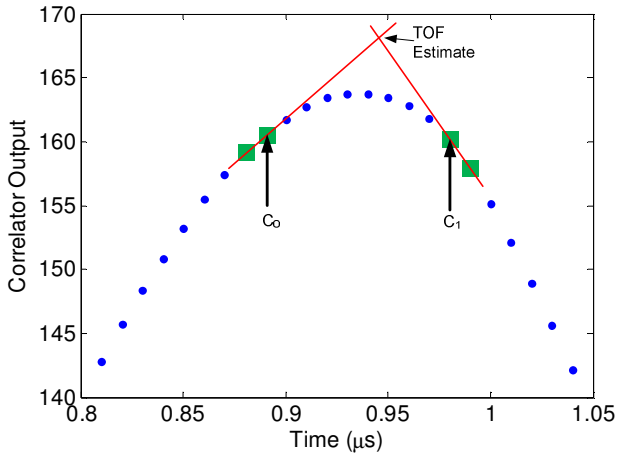


Figure 4.10 Plot showing the peak estimation method used in this system. First the locations of C_0 and C_1 are found such that the distance between C_0 and C_1 are fixed and the difference between the values at these points is near zero. The linear extrapolation from these points and neighboring points is used to find the peak estimate.

approximately equal in distance from the peak, and the line connecting the points on each side is found. The intersection point of the resulting lines is the peak location estimate, and it can be calculated to within the noise precision of the system. This method is best described graphically and is shown in Figure 4.10.

A total of 16 measurements are taken with one at each of 16 different center frequencies spread across the 2.4 GHz ISM band. Each of these measurements has the peak estimated in the manor described here, and the resulting 16 locations are the time of flight estimates. These estimates are then used together to produce a single range estimate that reduces the impact of multipath interference. In Chapter 3, we discussed how the demodulator is capable of making both positive and negative range estimate errors based on the relative phase of the direct and multipath signals. It was suggested that the best range estimate is to take the 25th percentile across the measurements from various carrier frequencies, and we do this calculation in Matlab as well and report this as the best case range estimate.

4.2.8 Localization Algorithm

The Waldo nodes combined together into a wireless network are intended for node localization. Although the localization process was not intended to be part of this work, a simple localization algorithm is implemented in Matlab. When ranges to three or more nodes with known location are estimated from one unknown location, the algorithm used provides an (x,y) location estimate. This algorithm is limited to this two-dimensional case for simplicity, but the ranging information could be used for full 3D localization with an appropriate localization algorithm. The algorithm implemented here is a simple least squares solution. The unknown location, L, is estimated to be the location that minimizes the squared error between the time of flight range estimates and the distances between L and the anchor nodes.

4.3 Range Measurement Cost

The cost of a range measurement can be measured in energy, time and hardware complexity. The energy cost of ranging when using the Waldo platform is very high because the platform is not low power compared to most embedded wireless devices. A prediction of energy costs based on available IC systems is a reasonable estimate of the energy cost of ranging using the algorithms presented here, and an estimate like this is included. The time and hardware complexity associated with this system is easier to quantify and will be discussed in this section as well.

A ranging operation consists of an RF segment and a computation segment. The time to complete each segment is shown in Table 4.5 assuming a 25 MHz clock rate for the processing. The ranging operation includes the time to switch to different carrier frequencies ($200\mu s$) and the time to change between transmit and receive mode ($16\mu s$), and these values are from the radio used in this implementation. The number of

operations to complete a correlation is calculated for an MSP430 microcontroller. The correlation is calculated for 50 of the 200 hundred data points (10^4 single sample correlations). The time for a ranging operation assumes the same parameters as used here: 48 copies transmitted of a 2 chip code at 1 Mchip/s. If the correlation is computed after each cycle, it would take 26 ms to complete all 16 ranging measurements. The typical coherence time of the indoor channels ranges from several tens of milliseconds to hundreds of milliseconds, so all of the measurements could be performed within the channel coherence time. The currently implemented system takes about 100 ms to complete all 16 range estimates, and this time is dominated by the time to transfer data to the PC.

In the current implementation, the power consumption for a ranging operation is significantly higher than would be the case in a more optimal system. The entire process is not carried out on the embedded system at this point because the correlation and other control is carried out on the PC. Therefore calculated energy numbers are not inherently meaningful. The energy numbers for just ranging are shown in Table 4.3, and these numbers are based on the time to perform each operation and the power numbers shown in Table 4.1. In the current system, the nodes idle in receive mode all the time (even when not ranging), and this dominates the overall power consumption (see Table 4.2).

Operation	**Base**	**Repeater**
Setup	0.681 mJ	0.681 mJ
Range	2.769 mJ	2.769 mJ
Total (16 channels)	3.450 mJ	3.406 mJ

Table 4.3 Table showing the energy cost of ranging on the Waldo platform as implemented. The setup cost includes the initial exchange of packets between the two nodes assuming the first packet transmission attempt is successful. The ranging cost is for a range estimation on a single channel, and it includes the cost of downloading data to the PC for analysis.

Function	Current	Power
Receive	19.7 mA	35 mW
Transmit	17.4 mA	31 mW
Processor	7.3 mA	13 mW

Table 4.4 Table showing the current and power for the radio and processor in different states. These numbers are based on the TI CC2420 and the TI MSP430, but the processor numbers are scaled to 25 MHz and 0.18 **µ**m process technology.

An estimation of the energy cost of ranging based on the power of currently available components provides insight into the energy cost of this algorithm in a complete system. These numbers are all scaled a $0.18\mu m$ CMOS to represent a single system-on-chip implementation. The Texas Instruments CC2420 IEEE 802.15.4 radio IC (designed in a $0.18\mu m$ CMOS) is used as a reference for radio power consumption, and the Texas Instruments MSP439F2617 is used as a reference for computation power consumption. The MSP430 (designed in a $0.35\mu m$ CMOS) power is scaled to be in a $0.18\mu m$ CMOS process running at 1.8V. The power associated with the radio and computation are shown in Table 4.4. Combining these power numbers with the time values, results in the energy costs shown in Table 4.5. The cost of sending or receiving a full packet using the CC2420 is about $200\mu J$ (using the numbers shown in Table 4.4). The energy cost of correlation is two-thirds of the total energy consumption because microprocessors are not the most efficient way to perform signal processing. A ranging operation is comparable to the cost of sending a few packets in this software solution.

A custom hardware implementation of the correlation would result in energy consumption less than that of a single packet transmission. Using the adder in [37] as a reference, we can find the energy required to perform a 16-bit addition for our reference process. We scale the power of this 32-bit adder implemented in $0.13\mu m$ CMOS at a 0.3V supply to a 16-bit adder implemented in $0.18\mu m$ CMOS at a 1.8V supply.

We also account for the change from 50MHz operation to 25MHz operation, and the cost of performing the correlation is calculated to be less than 100nJ (including the operation of the SRAM storing the data). The total cost of performing a full, 16 channel ranging operation under these assumptions is shown in Table 4.5 to be 170μJ. This is less than the cost of sending or receiving a packet (200 μJ) with the radio used for these calculations.

Single Frequency Operations	Time MSP430	Energy MSP430	Time ASIC	Energy ASIC
Ranging	112 μs	3.7 μJ	112 μs	3.7 μJ
Change frequency	200 μs	6.3 μJ	200 μs	6.3 μJ
Correlation time	1620 μs	21 μJ	420 μs	0.1 μJ

All 16 Ranging Operations	Time MSP430	Energy MSP430	Time ASIC	Energy ASIC
Total Time	31 ms	500 μJ	12 ms	170 μJ

Table 4.5 Table showing the time and energy to complete ranging operations for a solution using an MSP430 microcontroller and an application specific integrated circuit (ASIC) implementation. Power numbers used to calculate energy for the MSP430 solutions are based on those shown in Table 4.4. The ASIC solution is much more efficient in terms of energy, and it also completes the computations in less time.

Chapter 5

Ranging and Localization Demonstrations

The implemented system is capable of performing range measurements with accuracies meeting or exceeding those demonstrated by systems with greater instantaneous bandwidth and/or sampling rate [33, 38-40]. The key measurements described in this chapter are the noise performance as a function of SNR, ranging in both outdoor and indoor environments, and network localization.

5.1 Noise Performance

The noise performance of a ranging system should approach the Cramér-Rao bound on ranging performance. This suggests that variance of a range estimate in a white noise environment should decrease with increasing signal to noise ratio, and we should see performance better than the level set by the sampling rate. This can be tested experimentally, and this test is the subject of this section.

To measure the noise performance, two Waldo nodes are connected together via RF cables and a variable RF attenuator. This setup ideally ensures that only single path exists though the cable, and the measurement should not be impacted by multipath effects. In practice, however, the RF components on the Waldo boards are not shielded, and the radiation pattern of the boards provides a second path that not only results in multipath but also sets a limit to the maximum attenuation between the boards. One Waldo must be enclosed in a metal box to ensure this problem does not dominate the noise test. The power at the receiving antenna port was calibrated using a spectrum

analyzer. The SNR of the signal going into the baseband ADC was measured using an oscilloscope by measuring the noise floor with no signal and by measuring the signal amplitude when a signal is present. These measurements agreed well with the predicted signal to noise ratio based on the radio data sheet parameters. It is important to note that these Waldo nodes do not share a common time reference and the only thing connecting the two nodes together is an RF cable.

This test consisted of taking 100 range measurements at each attenuation setting and determining the standard deviation of these measurements. The demodulated data is sampled at 16 MHz yielding range bins of 19 m, and the signal bandwidth is 2 MHz. After demodulation, 32 copies of the 2 chip signal are averaged for a $t_s B$ product of 64 while maintaining code modulus synchronization. After the real time CMS operations are completed, the received signal is further analyzed in MATLAB as described in Chapter 4.

Figure 5.1 shows the standard deviation of ranging measurements as a function of baseband signal to noise ratio along with the Cramér-Rao bound for this system. At high values of SNR, the system does not achieve the Cramér-Rao bound because of the limited dynamic range of the system. In section 4.2.2 we discussed that the maximum SNR due to the dynamic range of the digital baseband is 18 dB. We expect that increasing SNR beyond 18dB at the low-IF will provide marginal improvements in ranging accuracy due to the limited dynamic range in the digital baseband, and this effect is observed in the figure. If the available dynamic range were larger, the measured noise performance would continue to improve with increasing SNR. Figure 5.1 also shows the bound associated with the range binning that would be observed based on the sampling rate of the system. It is clear that the measured noise performance exceeds

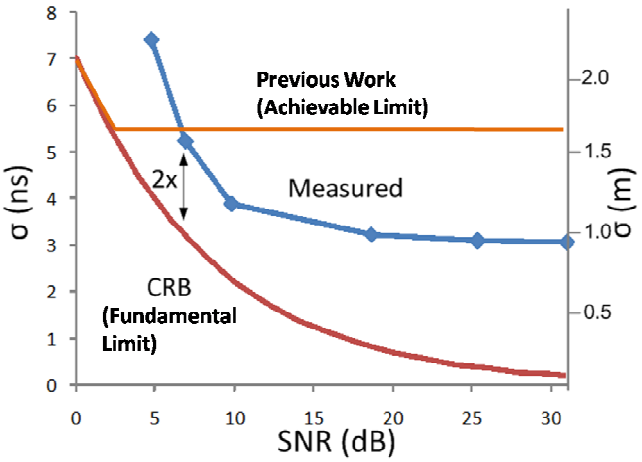


Figure 5.1 Measured noise performance of implemented Waldo system. The previous work achievable limit is due to range binning, and the fundamental limit is from the Cramér-Rao bound. The measured performance is within 2x of the Cramér-Rao bound until high SNR where the limited dynamic range of the digital baseband limits performance.

those possible in a range binning limited system. This system performs within a factor of 2 of the Cramér-Rao bound at SNRs of less than 10dB, and to achieve equivalent performance using two-way time transfer would have required a sample rate much higher than used here.

5.2 Outdoor Ranging Demonstration

Two Waldo nodes were used to perform ranging estimates in a parking lot with some cars but mostly open space. This environment provides a baseline for ranging performance in an environment where there is relatively little multipath interference. The two nodes are not connected together in any physical way, and the only method of communication is through the wireless link. The setup for ranging tests is shown in Figure 5.2. One node is mounted to a tripod with a battery pack, and the other node is on a cart tethered to the laptop with a USB cable. The ground truth distance between the nodes was measured using a tape measure. A range estimate was taken using the methods described in Chapters 3 and 4 at distances ranging from 1m to 45m, and the

Figure 5.2 Photographs of ranging setup showing both a fixed node on a tripod and the mobile node tethered to a laptop. The mobile node is the base node for ranging operations

received signal strength estimates are taken as well. The time of flight range estimates are shown in Figure 5.4. In order to produce this plot, the estimate produced using the algorithms described here is simply multiplied by the speed of flight to yield the slope of unity shown in the plot. The received signal strength (RSS) based range estimates are also shown in Figure 5.4. To generate these results, the equation transforming RSS to range that empirically minimized the mean squared ranging error was calculated based on the received signal power P_r and the wavelength, λ, of the signal used for the measurements.

$$d_{rss} = \left(\frac{0.1}{P_r}\right)^{\frac{2}{3}} \frac{\lambda}{5.41}$$

The result from free space was used as a guide to develop this expression.

$$d_{rss}|_{free\ space} = \left(\frac{P_t}{P_r}\right)^{1/2} \frac{\lambda}{4\pi}$$

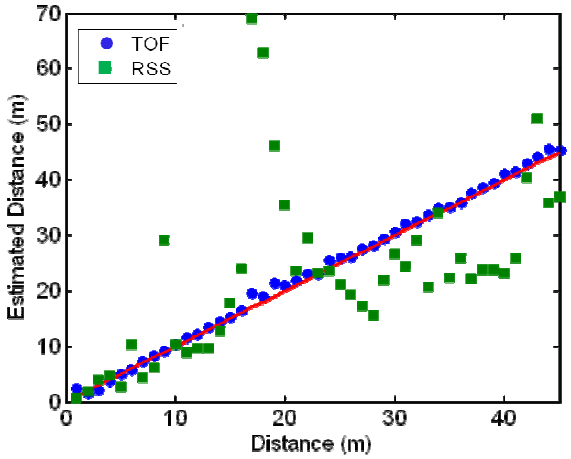


Figure 5.4 Plot showing outdoor time of flight ranging results where the red line is ground truth. These measurements were taken in a sparsely filled parking lot, and it is apparent that the received signal strength (RSS) range estimates did not provide good range estimates compared to the time of flight estimates. Even though this is a relatively mild multipath environment, the reflections that are present prevent RSS from providing reasonable accuracy, but the time of flight measurements perform much better.

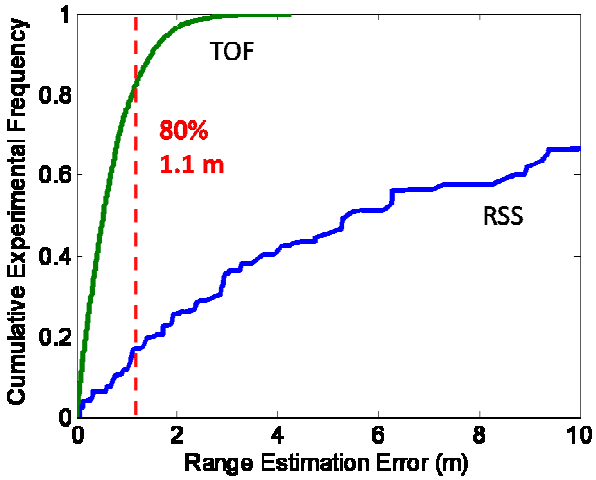


Figure 5.3 Experimental cumulative distribution of outdoor ranging measurements for both time of flight and received signal strength ranging methods.

The value for the transmitted power, P_t, is unknown, and it is commonly suggested that the square root may not be the best choice for many environments. This term along with the 4π constant and a possible exponent for the λ term were all considered when attempting to find the best conversion equation from RSS to range. These RSS range estimates were averaged across the 16 channels used to generate the results shown in Figure 5.4. The plotted RSS based results are the best that could be made using the data

Figure 5.5 Photo of hallway where ranging measurements shown in Figure 5.6 where taken.

recorded, and the results would have been worse if this error minimization had not been performed by comparing the ground truth to the range estimates. This sort of after the fact error minimization was not done for the time of flight estimates. Figure 5.3 shows the experimental cumulative distribution function for the time of flight ranging case and the received signal strength case. Approximately 80% of the time of flight measurements are accurate to within 1m, but not even 20% of the RSS based estimates are accurate to within 1m.

5.3 Indoor Ranging Demonstration

Indoor range estimates using the same setup as in described in section 5.2 have also been performed to verify that reasonable ranging accuracy can be achieved in environments typical to local area and sensor networks. The time of flight measurements shown in Figure 5.6 were taken in a hallway in Cory Hall on the University of California, Berkeley's campus, and the hallway environment is shown in Figure 5.5. The experimental cumulative distribution function of the ranging error is also shown in Figure 5.6. The achieved accuracy was better than 1 m 50% of the time and better than 3 m 80% of the time. There are no calibration steps or changes to the

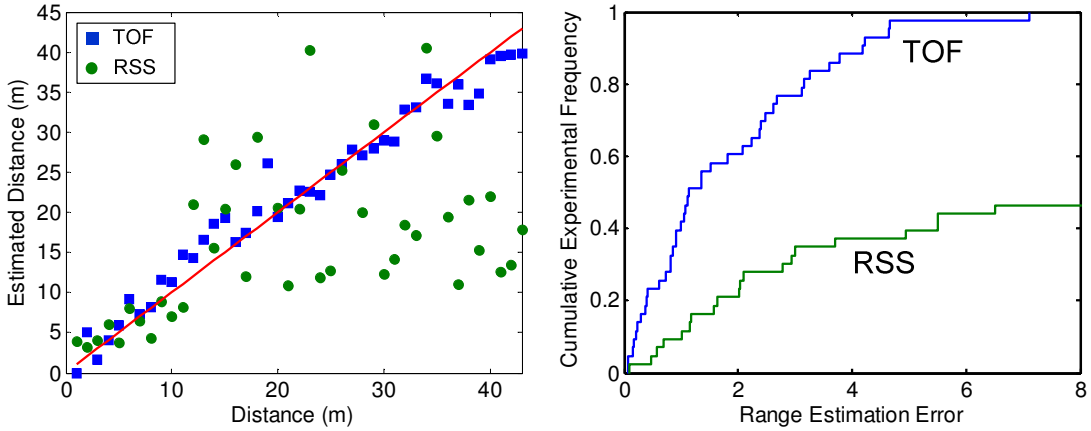


Figure 5.6 Plots showing ranging performance indoors for both time of flight and received signal strength range estimation. These measurements were taken in a hallway with several obstacles as shown in Figure 5.5. The time of flight estimates are significantly more consistent and accurate than the received signal strength measurements.

system firmware, software or calculation methods between this environment and the outdoor environment.

5.4 Localization Experiment

A simple localization experiment was performed to demonstrate the capability of Waldo to localize nodes using the implemented RF time of flight ranging system in a simple network. This experiment was carried out on a small open area between Evans Hall and

Figure 5.7 Photo showing the outdoor location where the localization experiment was conducted on an open space near Evans Hall on UC Berkeley's campus.

the Memorial Glade on UC Berkeley's campus. The approximate dimensions of the space are 50m by 40m of generally flat open space with some trees and bushes around the periphery. This environment is shown in Figure 5.7. Inter-node distances of up to 70m were available in this area, and communication and ranging could be performed at these distances without problems. Four static nodes were setup on tripods in the experiment space. The node tethered to the computer was held and carried through the field. Ground truth was measured using tape measures for both x and y position. The results of the localization experiment are shown in Figure 5.8. Localization accuracy is better than 2m for 80% of the estimates.

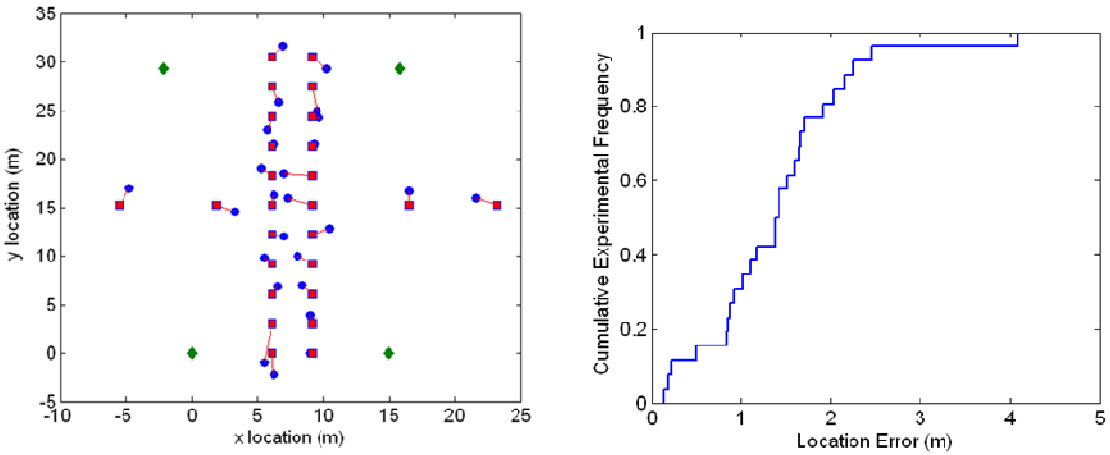


Figure 5.8 Left: Localization experiment results showing fixed reference nodes (green diamonds), ground truth locations (red squares) and estimated locations (blue circles). These measurements were taken in the location shown in Figure 5.7. Right: Experimental cumulative distribution of location accuracy from this experiment showing that 80% of measurements are accurate to better than 2m.

Chapter 6

Conclusions

6.1 Research Summary

Location aware wireless networks enable new applications such as local area asset management, ad-hoc deployments for the military, and node localization at network initialization. Current networks have either no or poor location awareness, and ultra-wideband technologies have appeared to be the only solution for location awareness. Developing narrowband methods for enabling location awareness will lead to much lower cost and more widely deployed networks.

This dissertation has presented a new method of performing round trip RF time of flight ranging that uses simple hardware and software to achieve 1-3m accuracy across a range of environments. The methods presented here achieve better noise performance than any other published two-way ranging system, and the multipath performance is equivalent or better than systems with wider bandwidth and more complicated multipath mitigation algorithms. Chapter 1 showed that location and ranging accuracy are comparable, and that accurate ranging is therefore critical to accurate localization. Chapter 2 presented the effects that limit ranging accuracy and some methods that are used to deal with these effects. Chapter 3 presented new methods for performing accurate ranging using limited hardware resources in difficult noise and multipath environments. Chapter 4 discussed the implementation of the Waldo platform, a wireless node that uses software defined radio to enable new

physical layers. The Waldo platform contains firmware implemented on an FPGA and embedded software that completes the implementation of the system. Chapter 5 presents experimental results from ranging and localization experiments demonstrating excellent performance.

6.2 Opportunities with Waldo

The Waldo platform is capable of being deployed in more advanced networks, ranging using other wireless baseband protocols, and testing other physical layer issues.

The network protocol implemented here is extremely simple, but the microprocessor used is the same as that used for most wireless sensor networks using TinyOS. A port of existing network protocols onto Waldo would enable many hour tests of location aware networks using the protocols presented here. Additional work must be done with Waldo to make this possible, however, and methods to do the correlation and range estimation would need to be included in either the firmware or embedded software.

The system implemented here is largely compatible with IEEE 802.15.4, but the platform could be used to demonstrate IEEE 802.11b (WiFi) or IEEE 802.15.1 (Bluetooth) ranging as well. The only changes required are in changing the baseband demodulation and modulation schemes. In the 802.15.1 case, all of the presented methods remain valid. The change to IEEE 802.11b, however, will also require a new multipath mitigation scheme. The scheme implemented here depends on using the low-IF demodulator, and a direct conversion receiver mode would need to be used for 802.11b.

A study of physical layer issues related to modulation schemes, fading performance, and physical layer error correction could all be implemented using Waldo.

These issues all require changes at the physical layer that would not be possible using a standard commercial radio, but they could be implemented in the firmware.

6.3 Ranging with Channel Estimation

The multipath mitigation strategy presented here does not rely on formal channel estimation, but the channel conditions are probed to improve accuracy. Acquiring sufficient channel knowledge to achieve comparable performance to that presented here is very challenging with the limited resources available. It seems possible, however, that the received signal strength profile across carrier frequency could be used to estimate some channel characteristics. A key advancement remaining is to find simple ways to estimate channel parameters to improve ranging performance. The key restrictions are that coherent measurements across carrier frequency are not possible using currently known techniques, and estimation methods must deal with this limitation.

REFERENCES

[1] D. Sexton, M. Mahony, M. Lapinski, and J. Werb, "Radio Channel Quality in Industrial Wireless Sensor Networks," in *Sensors for Industry Conference, 2005*, 2005, pp. 88-94.

[2] N. Patwari, J. N. Ash, S. Kyperountas, A. O. Hero, III, R. L. Moses, and N. S. Correal, "Locating the nodes: cooperative localization in wireless sensor networks," *Signal Processing Magazine, IEEE,* vol. 22, pp. 54-69, 2005.

[3] G. Mao and B. Fidan, "Localization Algorithms and Strategies for Wireless Sensor Networks," Information Science Reference, 2009, p. 526.

[4] S. Lanzisera and K. S. J. Pister, "RF Ranging Methods and Performance Limits for Sensor Localization," in *Localization Algorithms and Strategies for Wireless Sensor Networks*, G. Mao and B. Fidan, Eds.: Information Science Reference, 2009, p. 526.

[5] "IEEE Standard for Information technology- Telecommunications and information exchange between systems- Local and metropolitan area networks- Specific requirements Part 15.4: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Low-Rate Wireless Personal Area Networks (WPANs)," *IEEE Std 802.15.4-2006 (Revision of IEEE Std 802.15.4-2003),* pp. 0_1-305, 2006.

[6] A. J. Weiss and J. S. Picard, "Maximum Likelihood localization of wireless networks using biased range measurements," in *Communications and Information Technologies, 2007. ISCIT '07. International Symposium on*, 2007, pp. 865-870.

[7] K. Xinghong and S. Huihe, "Maximum Likelihood Localization Algorithm Using Wireless Sensor Networks," in *Innovative Computing, Information and Control, 2006. ICICIC '06. First International Conference on*, 2006, pp. 263-266.

[8] J. Sheinvald, M. Wax, and A. J. Weiss, "On maximum-likelihood localization of coherent signals," *Signal Processing, IEEE Transactions on,* vol. 44, pp. 2475-2482, 1996.

[9] I. Ziskind and M. Wax, "Maximum likelihood localization of multiple sources by alternating projection," *Acoustics, Speech and Signal Processing, IEEE Transactions on,* vol. 36, pp. 1553-1560, 1988.

[10] M. W. Carter, H. H. Jin, M. A. Saunders, and Y. Ye, "SpaseLoc: An Adaptive Subproblem Algorithm for Scalable Wireless Sensor Network Localization," *SIAM J. on Optimization,* vol. 17, pp. 1102-1128, 2006.

[11] F. Anjum, S. Pandey, and P. Agrawal, "Secure localization in sensor networks using transmission range variation," in *Mobile Adhoc and Sensor Systems Conference, 2005. IEEE International Conference on*, 2005, pp. 9 pp.-203.

[12] H. L. Van Trees, *Detection, estimation, and modulation theory*. New York: Wiley, 2001.

[13] E. D. Kaplan, *Understanding GPS : principles and applications*. Boston: Artech House, 1996.

[14] S. Lanzisera, D. T. Lin, and K. S. J. Pister, "RF Time of Flight Ranging for Wireless Sensor Network Localization," in *Intelligent Solutions in Embedded Systems, 2006 International Workshop on*, 2006, pp. 1-12.

[15] B. D. D. Lachartre, D. Morche, L. Ouvry, M. Pezzin, B. Piaget, J. Prouvée, P. Vincent, "A 1.1nJ/b 802.15.4a-Compliant Fully Integrated UWB Transceiver in 0.13µm CMOS," in *International Solid State Circuits Conference (ISSCC)* San Francisco, CA, 2009.

[16] L. Doherty, W. Lindsay, and J. Simon, "Channel-Specific Wireless Sensor Network Path Data," in *Computer Communications and Networks, 2007. ICCCN 2007. Proceedings of 16th International Conference on*, 2007, pp. 89-94.

[17] S. Lanzisera, A. Mehta, and K. S. J. Pister, "Reducing Average Power in Wireless Sensor Networks Through Data Rate Adaptation," in *IEEE International Conference on Communication (ICC)* Dresden, Germany: IEEE, 2009.

[18] T. Williamson and N. A. Spencer, "Development and operation of the Traffic Alert and Collision Avoidance System (TCAS)," *Proceedings of the IEEE,* vol. 77, pp. 1735-1744, 1989.

[19] D. Kirchner, "Two-way time transfer via communication satellites," *Proceedings of the IEEE,* vol. 79, pp. 983-990, 1991.

[20] M. A. Richards, *Fundamentals of radar signal processing*. New York: McGraw-Hill, 2005.

[21] P. G. Hoel, *Introduction to mathematical statistics*, 5th ed. New York: Wiley, 1984.

[22] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-time signal processing*, 2nd ed. Upper Saddle River, N.J.: Prentice Hall, 1999.

[23] Q. H. Spencer, B. D. Jeffs, M. A. Jensen, and A. L. Swindlehurst, "Modeling the statistical time and angle of arrival characteristics of an indoor multipath

channel," *Selected Areas in Communications, IEEE Journal on,* vol. 18, pp. 347-360, 2000.

[24] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge, UK ; New York: Cambridge University Press, 2005.

[25] S. Shah and A. Tewfix, "Enhanced Position Location with UWB in Obstructed LOS and NLOS Multipath Environmnets," in *XIII European Signal Processing Conference*, 2005.

[26] N. Udar, K. Kant, R. Viswanathan, and D. Cheung, "Ultra Wideband Channel Characterization and Ranging in Data Centers," in *Ultra-Wideband, 2007. ICUWB 2007. IEEE International Conference on*, 2007, pp. 322-327.

[27] F. M. Dickey, L. A. Romero, and A. W. Doerry, "Superresolution and Synthetic Aperture Radar," S. N. Laboratories, Ed. Albuquerque, New Mexico: Sandia National Laboratories, 2001.

[28] M. A. Pallas and G. Jourdain, "Active high resolution time delay estimation for large BT signals," *Signal Processing, IEEE Transactions on,* vol. 39, pp. 781-788, 1991.

[29] N. Dharamdial, R. Adve, and R. Farha, "Multipath delay estimations using matrix pencil," in *Wireless Communications and Networking, 2003. WCNC 2003. 2003 IEEE*, 2003, pp. 632-635 vol.1.

[30] N. Nefedov and M. Pukkila, "Iterative channel estimation for GPRS," in *Personal, Indoor and Mobile Radio Communications, 2000. PIMRC 2000. The 11th IEEE International Symposium on*, 2000, pp. 999-1003 vol.2.

[31] K. Pahlavan, L. Xinrong, and J. P. Makela, "Indoor geolocation science and technology," *Communications Magazine, IEEE,* vol. 40, pp. 112-118, 2002.

[32] L. Song, R. Adve, and D. Hatzinakos, "Matrix Pencil for Positioning in Wireless ad hoc Sensor Network," in *Wireless Sensor Networks*, 2004, pp. 18-27.

[33] S. Lanzisera and K. S. J. Pister, "Burst Mode Two-Way Ranging with Cramer-Rao Bound Noise Performance," in *Global Telecommunications Conference, 2008. IEEE GLOBECOM 2008. IEEE*, 2008, pp. 1-5.

[34] K. Mandke, C. Soon-Hyeok, K. Gibeom, R. Grant, R. C. Daniels, K. Wonsoo, R. W. Heath, and S. M. Nettles, "Early Results on Hydra: A Flexible MAC/PHY Multihop Testbed," in *Vehicular Technology Conference, 2007. VTC2007-Spring. IEEE 65th*, 2007, pp. 1896-1900.

[35] J. Dunlop, *Telecommunications engineering*, 3rd ed. New York: Chapman & Hall, 1994.

[36] R. E. Phelts, "Multicorrelator Techniques for Robust Mitigation of Threats to GPS Signal Quality," in *Mechanical Engineering*. vol. Ph.D. Stanford, CA: Stanford University, 2001.

[37] K. Ishibashi, T. Yamashita, Y. Arima, I. Minematsu, and T. Fujimoto, "A 9μW 50MHz 32b adder using a self-adjusted forward body bias in SoCs," in *Solid-State Circuits Conference, 2003. Digest of Technical Papers. ISSCC. 2003 IEEE International*, 2003, pp. 116-482 vol.1.

[38] C. Hoene and J. Willmann, "Four-way TOA and software-based trilateration of IEEE 802.11 devices," in *Personal, Indoor and Mobile Radio Communications, 2008. PIMRC 2008. IEEE 19th International Symposium on*, 2008, pp. 1-6.

[39] T. C. Karalar and J. Rabaey, "An RF ToF Based Ranging Implementation for Sensor Networks," in *Communications, 2006. ICC '06. IEEE International Conference on*, 2006, pp. 3347-3352.

[40] S. Schwarzer, M. Vossiek, M. Pichler, and A. Stelzer, "Precise distance measurement with IEEE 802.15.4 (ZigBee) devices," in *Radio and Wireless Symposium, 2008 IEEE*, 2008, pp. 779-782.