August 27, 2002

Algorithms for GPS Operation Indoors and Downtown

Nainesh Agarwal\(\alpha\), Julien Basch\(\alpha\), Paul Beckmann\(\alpha\), Piyush Bharti\(\alpha\), Scott Bloebaum\(\alpha\)
Stefano Casadei\(\alpha\), Andrew Chou\(\beta\), Per Enge\(\alpha\)\(\beta\), Wungkum Fong\(\alpha\), Neesha Hathi\(\alpha\),
Wallace Mann\(\alpha\), Anant Sahai\(\gamma\), Jesse Stone\(\alpha\), John Tsitsiklis\(\alpha\)\(\delta\), Benjamin Van Rooy\(\alpha\)\(\beta\)

\(\alpha\)Enuvus, Inc.
\(\beta\)Stanford University
\(\gamma\)University of California, Berkeley
\(\delta\)Massachusetts Institute of Technology

Abstract

The proliferation of mobile devices and the emergence of wireless location-based services has generated consumer demand for availability of GPS in urban and indoor environments. This demand calls for enhanced GPS algorithms that accommodate high degrees of signal attenuation and multipath effects unique to the “urban channel.” This paper overviews the market for location-based services and discusses algorithmic innovations that address challenges posed by urban environments.
1 Introduction

The proliferation of mobile devices has generated consumer demand for location-based services. With this emerging market comes new performance requirements for GPS. Unlike the military and vehicle navigation applications of the past, to support location-based services, GPS must perform robustly in urban settings – indoors and outdoors. Receivers must be small and inexpensive and must compute location in seconds. Conventional GPS receivers fall short because they are not equipped to deal with the degree of attenuation and multipath present in urban environments and take too much time to generate a location fix.

Two advances in GPS technology are required to meet these new challenges:

1. **Assistance Data.** This is information useful to reception and analysis of a GPS signal. Examples include a time stamp, the location of a nearby base station, ephemeris information, and navigation data. In a transcendent discovery of the 1980's, NASA engineers Ralph E. Taylor and James W. Sennott observed that use of such data can enhance receiver sensitivity and dramatically reduce time-to-first-fix. These engineers designed the Assisted GPS (A-GPS) architecture (Taylor and Sennott, 1984), in which assistance data is transmitted from a wireless base station to a GPS receiver in order to enhance performance. Unfortunately, during the 1980’s and early 1990’s, it was not practical to transmit data to mobile consumer devices. The recent widespread adoption of wireless networks has made transmission of assistance data a reality.

2. **Advances in GPS Signal Processing Algorithms.** One approach to A-GPS involves porting conventional GPS algorithms to an architecture that supports the transmission of assistance data. While the performance of conventional algorithms will benefit from the availability of assistance data, this performance still falls far short of consumer expectations, especially in urban environments. In order to effectively support location-based services, GPS signal processing algorithms need to be redesigned from the ground up, specifically to address challenges posed by urban environments.

This paper discusses some key algorithmic ideas required to achieve the sensitivity, time-to-first-fix, and accuracy demands of location-based services. We review relevant literature and also discuss a number of innovations that have grown out of the development of UrbanGPS™ – a new generation of GPS algorithms designed at Envisus. Our discussion of algorithms will focus on the following thrusts:

1. **Coherent Processing.** Conventional GPS algorithms process blocks of one or several milliseconds of signal coherently and combine the results noncoherently. Receiver sensitivity can be amplified if longer durations of signal – e.g., one or more seconds – are processed coherently. However, coherent processing over long durations introduces new technical challenges that call for significant algorithmic innovations.

2. **Information Fusion.** Conventional GPS algorithms acquire signals from different satellites in a decoupled manner. This treatment ignores the fact that acquisition of one or more satellite signals can assist in the acquisition of additional signals. Fusion of information from multiple satellites can enhance receiver sensitivity.
3. **Anti-Multipath Triangulation.** Multipath is the dominant source of error in urban environments. Conventional methods of multipath mitigation, designed primarily for open-sky applications, do not effectively accommodate the urban setting. Anti-multipath triangulation leverages statistical models based on data collected in urban environments to reduce errors.

4. **Time Stamp Recovery.** A critical piece of information available to algorithms operating in an A-GPS architecture is the time stamp. In a CDMA cellular network, base stations are typically able to provide time stamps that are accurate to within tens of microseconds, and this simplifies the design of A-GPS algorithms. In some other networks, time stamps can err by a second or more. Novel algorithms are required to accommodate such large time stamp errors.

This paper is organized as follows. In the next section, we discuss market requirements for location-based services and the challenges they pose to GPS. Section 3 describes the role of A-GPS algorithms. Subsequent sections discuss various innovations required to overcome technical challenges introduced by the new market.

## 2 Market Demand and Challenges

Performance demands on GPS are being driven by the market for location-based services (LBS). In the next subsection, we discuss the LBS market, representative applications, and the requirements they impose on location technology. Most of the high-value LBS applications call for the high levels of accuracy that can be offered by GPS. In Section 2.2 we discuss the challenges that the LBS market poses to GPS.

### 2.1 The Market for Location-Based Services

The LBS market is projected to grow dramatically over the next several years. Analysts predict annual revenue to be anywhere from $18 billion (Taylor, 2001) to $20 billion (Green and Betti, 2000) to $33 billion (Saunders et al., 2001) by 2006. To date, over 50 wireless carriers in Europe and Asia have deployed LBS, generally using low accuracy technology (e.g., cell ID technology, with a location uncertainty of 300 meters to 2 kilometers). Due to the low level of precision available, carriers can only support a limited number of applications at this time. LBS applications of higher value to consumers tend to require better accuracy.

There are five categories of location-based services:

1. **Information Services.** This category includes enhanced directory assistance, traffic information, and navigation services. Although some information service applications can function with low accuracy location technology, much of the value to the consumer will come with increased accuracy.

---

1We use the term *triangulation* loosely. Though the correct term in this context would be *trilateration*, use of the term *triangulation* has become the norm in the GPS industry.
2. **Trigger services.** Location sensitive billing and event-based advertising and promotions are applications that are classified as trigger services. Location sensitive billing allows wireless carriers to compete with landline connectivity providers. Event-based advertising can help a store to attract potential customers in the vicinity.

3. **Entertainment.** A large portion of entertainment related LBS consists of gaming applications. Location-enabled games are anticipated to be multi-player, involving small groups and/or a mass market.

4. **Safety.** In the U.S., safety is the primary market driver for location technology deployment. While the E-911 mandate in the U.S. and similar anticipated regulation in the European and East Asian markets is the most recognizable portion of the safety application segment, it is not expected to derive substantial revenues for wireless operators. Revenues will more likely come from applications such as roadside assistance and personal security.

5. **Third Party Tracking.** This consists of enterprise applications, such as fleet management and the tracking of assets, and consumer applications, primarily of the people-finder type. Services to track children and the elderly have already been launched in some regions. People finder applications are also gaining popularity, particularly among younger consumers.

The table below summarizes requirements associated with various applications mentioned above. There is clearly strong demand for the high levels of accuracy that can be offered by GPS.

<table>
<thead>
<tr>
<th>Application</th>
<th>Coverage</th>
<th>Accuracy</th>
<th>Time-to-First-Fix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enhanced Directory Assistance</td>
<td>indoor, urban, rural</td>
<td>25-100m</td>
<td>5-15s</td>
</tr>
<tr>
<td>Navigation</td>
<td>urban, rural</td>
<td>1-25m</td>
<td>1-5s</td>
</tr>
<tr>
<td>Traffic Information</td>
<td>urban, rural</td>
<td>300m</td>
<td>5-15s</td>
</tr>
<tr>
<td>Location Sensitive Billing</td>
<td>indoor, urban, rural</td>
<td>5-100m</td>
<td>1-5s</td>
</tr>
<tr>
<td>Event-Based Advertising</td>
<td>indoor, urban, rural</td>
<td>50-125m</td>
<td>5-15s</td>
</tr>
<tr>
<td>Gaming</td>
<td>indoor, urban, rural</td>
<td>5-250m</td>
<td>1-15s</td>
</tr>
<tr>
<td>Emergency Services</td>
<td>indoor, urban, rural</td>
<td>50-100m</td>
<td>5-30s</td>
</tr>
<tr>
<td>Roadside Assistance</td>
<td>urban, rural</td>
<td>75-125m</td>
<td>5-30s</td>
</tr>
<tr>
<td>Personal Security</td>
<td>indoor, urban, rural</td>
<td>1-50m</td>
<td>5-15s</td>
</tr>
<tr>
<td>Person/Asset Tracking</td>
<td>indoor, urban, rural</td>
<td>30-200m</td>
<td>1-30s</td>
</tr>
<tr>
<td>People Finder</td>
<td>indoor, urban</td>
<td>1-100m</td>
<td>5-15s</td>
</tr>
</tbody>
</table>

### 2.2 The Urban Challenge

Key GPS algorithm performance metrics relevant to LBS applications include:

- **Sensitivity.** The level of $C/N_0$ in dB-Hz required for reliable acquisition of a satellite signal. We take $C/N_0$ to denote the signal power divided by noise power spectral density, measured at the input to an A/D converter.²

²Engineers sometimes alternatively define sensitivity in terms of signal power requirements at an antenna. Given antenna and RF front end characteristics and assumptions on noise present in the environment, this signal power can be mapped to $C/N_0$. In order to focus in this paper on algorithm performance, as opposed
Figure 1: Histograms of $C/N_0$ for signals acquired in open sky, in an outdoor urban environment in Tokyo, and in a Tokyo mall.

- **Accuracy.** The root-mean-squared error in meters among location fixes that are not false alarms.
- **Time-to-First-Fix (TTFF).** The time in seconds required to compute a position fix assuming cold-start conditions (i.e., no information from previous fixes).

Applications discussed in the previous subsection call for a location technology that offers robust coverage in rural and urban environments – indoors and outdoors – and this drives sensitivity requirements. Many of the applications require 50m or better accuracy, and there is utility associated with improved accuracy down to the meter level. Furthermore, this coverage and accuracy should be delivered with a TTFF of less than ten seconds. In this section, we present some of the challenges that had to be addressed in developing a technology to meet these demands. In particular, we discuss the levels of sensitivity required to operate in urban environments and the degree of multipath that must be mitigated to offer desired accuracy.

There has been some work in the GPS literature to assess and find ways of dealing with attenuation caused by foliage (Spilker, 1996c), but virtually none on the effects of urban environments. Figure 1 speaks to the degradation in $C/N_0$ that arises in urban environments. Each plot was produced based on GPS signals acquired at a particular location. Each presents three histograms – one of $C/N_0$ for the strongest satellite signal acquired in each signal capture, one for $C/N_0$ of the second strongest signal acquired, and one for $C/N_0$ of the remaining acquired signals. The first plot is based on data collected under open-sky conditions. Note that the strongest acquired signal is almost always at or above 50dB-Hz.

The second and third plots were generated using data collected at an outdoor urban location and inside an urban mall, respectively, both in downtown Tokyo. Observe that the $C/N_0$ values are substantially lower than in an open sky environment. In an outdoor urban environment, signals are typically attenuated by 10 to 25dB, while indoors in the city, signals are typically attenuated by 25 to 35dB. These degrees of attenuation are well beyond what conventional GPS algorithms can accommodate.

Urban multipath presents another major hurdle. Figure 2 displays two bar graphs. The first represents a signal capture in an open-sky environment, and the second a signal capture to receiver performance, we define sensitivity in terms of $C/N_0$.  


in an outdoor location in Tokyo from which the photograph in the figure was taken. Each bar graph plots pseudorange errors associated with signals from overhead satellites. As seen in the first bar graph, in a clear sky environment, pseudorange errors are typically less than a few meters. The second bar graph shows how pseudorange errors can be erratic and large in an urban setting. Note that three of the pseudoranges err by more than 300m. Such large errors translate to poor accuracy if conventional GPS algorithms are employed.

3 Basics of A-GPS Algorithms

We will discuss later in this paper advances that enable effective GPS operation despite the aforementioned challenges. But first, let us present in this section some background on A-GPS and the general make up of an A-GPS algorithm.

In the A-GPS architecture, in addition to a digitized GPS signal, several pieces of assistance data are made available to the GPS algorithm, including:

1. Time stamp. This can be supplied through a cellular network and represents an estimate of the time at which the GPS signal capture was initiated. In a CDMA network, time stamps are typically accurate to within 100µs or better. In a GSM network, time stamps can be off by several seconds.

2. Approximate location. Typically taken to be the location of the base station from which the mobile device receives assistance data, the approximate location serves as a coarse estimate of the receiver’s location. In urban areas, the closest base station is typically within a few kilometers of the receiver. In rural areas, the closest base station can be tens of kilometers from the receiver.

3. Ephemeris information. This is easily obtained through a network, and can be used to compute satellite locations, velocity, and acceleration.

4. Satellite clock corrections. Satellite clocks drift over time. At any given time, clock error estimates can be obtained through the network.

5. Differential corrections. As with conventional differential GPS systems, this data is obtained from a reference receiver network and enhances system accuracy.
6. **Navigation data.** Navigation data is required for coherent processing of long durations of signal. With the right algorithms, transmission of navigation data from the base station to the mobile device can greatly enhance sensitivity.

Current standards defined for CDMA and GSM networks support transmission of the assistance data enumerated above or equivalents.

The job of an A-GPS algorithm is to estimate receiver location based on assistance data and the received GPS signal. An important aspect of A-GPS is that there is no need for decoding navigation data. This is important because:

1. Conventional GPS receivers acquire signals, then must gather and decode navigation data for at least eighteen seconds (the duration of the first three subframes) and generally over thirty seconds prior to generating a location fix. Since A-GPS algorithms do not need to decode navigation data, it is possible to reduce the time-to-first-fix dramatically.

2. The bit-error-rate associated with decoding navigation data increases quickly as signals are attenuated by 10 to 20dB. This precludes the high degrees of sensitivity.

Conventional receivers decode navigation data to obtain several important pieces of information, including the signal transmission time, ephemeris information, satellite clock corrections, and ionospheric delay corrections. We refer the reader to (Misra and Enge, 2001) for a more detailed account of navigation data and how it is decoded and used in conventional receivers. In A-GPS, the last three items do not need to be derived from navigation data – ephemeris information and satellite clock corrections are provided as assistance data while the ionospheric delay corrections are superseeded by differential corrections. In conventional GPS algorithms, signal transmission times are critical to pseudorange calculations. However, given a time stamp and approximate location, A-GPS algorithms can determine pseudoranges without access to signal transmission times.

For concreteness, let us describe at a high-level steps of an A-GPS algorithm. These steps are based on ideas that have appeared in the A-GPS literature (e.g., Taylor and Sennott, 1984), which build on more traditional GPS algorithms (see, e.g., Misra and Enge, 2001).

We begin by introducing notation and simplifying assumptions that will keep the exposition brief. Let \( \ell \) denote the receiver’s location and \( \tilde{\ell} \) the approximate (i.e., base station) location in Earth-centered Earth-fixed (ECEF) coordinates. Assume that \( \| \ell - \tilde{\ell} \|_2 \leq 10 \text{km} \), where \( \| \cdot \|_2 \) denotes distance. Assume that there is no rotation undergone by the Earth during the time period of interest. Let \( t \) be the time at the start of signal capture and \( \tilde{t} \) the time stamp. We will assume that the time stamp error \( |t - \tilde{t}| \) is less than 100\( \mu \text{s} \) (we discuss in Section 7 methods of dealing with a less accurate time stamp). Let us index satellites visible to the receiver by \( 1, \ldots, N \). We assume that \( N \geq 4 \) and that the set of satellites visible from the receiver’s location \( \ell \) is the same as that visible from the approximate location \( \tilde{\ell} \).

Let \( s_i \) be the position in ECEF coordinates of the \( i \)th satellite at the time of transmitting the signal captured at time \( t \). The range \( r_i \) for satellite \( i \) is defined by \( r_i = \| s_i - \ell \|_2 \). Let us assume that there are no significant atmospheric delays, so that the range satisfies \( r_i = (t - t_i) c \), where \( t_i \) is the time at which the signal received at time \( t \) was transmitted and \( c \) is the speed of light. We will measure code phase in units of time in the range \([0, 1] \text{ms}\).
Without loss of generality, assume that the code phase at time 0 is 0. Let $\tau_i$ be the code phase received at time $t$. The time $t_i$ that the signal received at time $t$ was transmitted is then given by $t_i = K_i 10^{-3} + \tau_i$, for some integer $K_i$. Solving for $K_i$ in terms of $t$ and $\tau_i$, we get $K_i = \left( \frac{(t - r_i/c - \tau_i)}{10^3} \right)$. Let $\hat{\tau}_i$ denote the range of the $i$th satellite, given by $\hat{\tau}_i = \|s_i - \hat{t}\|_2$. Note that $K_i = \left( \frac{(\hat{t} - \hat{\tau}_i/c - \hat{\tau}_i)}{10^3} \right)$, where the brackets denote rounding to the nearest integer and $\hat{\tau}_i$ denotes an estimate of $\tau_i$, if $|\left( \frac{(\hat{t} - \hat{\tau}_i/c - \hat{\tau}_i)}{10^3} - (t - r_i/c - \tau_i) \right)| < 0.5$. This is easily satisfied, for example, if time stamp error is less than 100µs and approximate location error is less than 10km, as assumed, and the error on the code phase estimate $\hat{\tau}$ is less than 1km.

The main steps of the algorithm are summarized as follows:

1. **Compute satellite positions and identify visible satellites.** Based on the ephemeris information and the time stamp $\hat{t}$, compute positions of all satellites. Identify the collection of satellites 1, . . . , $N$ visible from $\hat{t}$.

2. **Acquire signals.** For each satellite 1, . . . , $N$, acquire the signal to estimate the code phase. Denote estimated code phases by $\hat{\tau}_1, \ldots, \hat{\tau}_N$.

3. **Compute pseudoranges.** For each $i$, compute $\hat{K}_i = \left( \frac{(\hat{\tau}_i/\hat{t} - \hat{\tau}_i)}{10^3} \right)$. Then, let $\hat{t}_i = K_i 10^{-3} + \hat{\tau}_i$, and compute pseudoranges $\rho_i = (\hat{t}_i - t_i) c$.

4. **Triangulate.** Note that there are two sources of error in $\rho_i$. First, there is the error $\hat{\tau}_i - \tau_i$ in the code phase estimate. This error is well modeled as a zero-mean Gaussian random variable with some variance $\sigma_i^2$ that can be estimated in the code phase acquisition process. Second, there is the time stamp error $\hat{t}_i - \tau_i$, which is satellite-independent. This term can be corrected during least-squares triangulation. In particular, to compute location and the time stamp error $\delta = \hat{t} - t$, we solve the nonlinear least-squares problem

$$
\min_{\ell \in \mathbb{R}^3, \delta \in \mathbb{R}} \sum_{i=1}^{N} \frac{1}{\sigma_i^2} (\rho_i - \delta - \|\hat{t}_i - s_i\|_2)^2.
$$

This problem can be addressed by any of a number of computational methods (see, e.g., Misra and Enge, 2001; Bancroft, 1985).

The simple algorithm described above benefits greatly from the availability of assistance data. In particular, it does not need to decode navigation data. This significantly improves time-to-first-fix and sensitivity. In particular, even in conventional GPS receivers, signal acquisition is reliable at lower levels of $C/\tilde{N}_0$ than is navigation data decoding.

Despite the advantages afforded by assistance data, use of conventional algorithms for signal acquisition and triangulation (steps 2 and 4 above) in the A-GPS architecture does not adequately address the urban challenge. Substantial further improvements in sensitivity are required. Furthermore, least-squares triangulation methods — even if used in conjunction with conventional multipath mitigation algorithms — yield very poor accuracy in urban environments. The following sections discuss technological advances that enable A-GPS to accommodate requirements of the LBS market.
4 Coherent Processing

Conventional GPS algorithms coherently process blocks of GPS signal spanning anywhere from one to twenty milliseconds in duration. The results are then combined noncoherently — that is, the squares or magnitudes of block correlation functions are summed. This noncoherent combination incurs “squaring loss,” reducing sensitivity (Misra and Enge, 2001). Squaring loss is avoided if the receiver coherently processes over the entire duration of signal captured (e.g., one to three seconds). However, several obstacles prevent coherent processing in conventional receivers:

1. Coherent integration of more than 20ms of GPS signal requires knowledge of the satellite’s navigation message.
2. Conventional algorithms for coherent processing impose onerous computational requirements.
3. Coherent integration over long durations requires sophisticated signal models to take into account a number of phenomena that are typically ignored.

The first obstacle is lifted by presence navigation messages in assistance data. Advances in algorithms and modeling techniques enable coherent integration over multiple seconds.

In the following subsection, we discuss coherent and noncoherent processing and performance benefits afforded by the former. Subsection 4.2 introduces computational challenges and methods associated with coherent processing. The need for sophisticated models and further computational challenges associated with such models are discussed in Subsection 4.3.

4.1 Coherent and Noncoherent Processing

Let us consider the problem of acquiring a GPS signal from a single satellite. We begin with a received signal, which we will assume has been preprocessed by down-conversion to in-phase and quadrature baseband components (the down-conversion is based on the GPS carrier frequency plus satellite Doppler estimated from assistance data). We will also assume for simplicity that there is no navigation data. Hence, the signal can be modeled as

$$x(t) = \alpha y(t - \tau)e^{-2\pi j(f(t + \phi))} + w(t),$$

where $\alpha^2$ is the $C/N_0$, $y$ is the repeated PRN sequence of period 1ms, $\tau$ is the code phase, $f$ is the residual Doppler caused by receiver clock instabilities, user velocity, and differences between the approximate and actual receiver location, $\phi$ is the carrier phase, and $w$ is complex white Gaussian noise with unit power.

Assistance information provides intervals $[\tau, \tau']$ and $[f, f']$ that are known to contain the code phase and Doppler. The problem of acquiring a GPS signal is generally formulated in terms of maximum likelihood estimation over $\tau \in [\tau, \tau']$, $f \in [f, f']$, and $\alpha e^{-2\pi j\phi} \in \mathcal{C}$ (the complex amplitude). Given a signal captured from time 0 to $T$, estimates of $\tau$ and $f$ are given by:

$$\arg\max_{\tau \in [\tau, \tau']} A_T(\hat{\tau}, \hat{f})$$

8
where the coherent ambiguity function $A_T$ is defined by

$$A_T(\hat{\tau}, \hat{f}) = \left| \int_{t=0}^{T} x(t) y(t - \hat{\tau}) e^{2\pi j \hat{f} t} dt \right|^2. \tag{2}$$

Coherent processing simply refers to solution of the above problem. As we will discuss in the next section, simple approaches to solving this problem become computationally prohibitive as the duration $T$ of signal capture grows. Noncoherent processing is an approximation method that reduces the computational burden. Noncoherent processing involves partitioning the received signal into $K$ blocks of size $b = T/K$, and estimating code phase and Doppler by maximizing a noncoherent ambiguity function given by either

$$A_T(\hat{\tau}, \hat{f}) = \left( \sum_{k=1}^{K} \left| \int_{t=(k-1)b}^{kb} x(t) y(t - \hat{\tau}) e^{2\pi j \hat{f} t} dt \right| \right)^2, \tag{3}$$

or

$$A_T(\hat{\tau}, \hat{f}) = \sum_{k=1}^{K} \left| \int_{t=(k-1)b}^{kb} x(t) y(t - \hat{\tau}) e^{2\pi j \hat{f} t} dt \right|^2. \tag{4}$$

The first (Equation 3) corresponds to solving a maximum likelihood problem in which $\alpha$ is assumed to remain constant over the entire duration, while $\phi$ is allowed to depend on the block index $k$, changing from block to block. The second (Equation 4) corresponds to solving a maximum likelihood problem in which both $\alpha$ and $\phi$ are allowed to depend on $k$. Equation (3) leads to slightly better sensitivity than Equation (4), assuming that $\alpha$ is indeed constant.

The primary objective of signal acquisition is to obtain a good estimate of the code phase. An acquisition is said to be false if the resulting code phase estimate $\hat{\tau}$ differs by more than one microsecond (i.e., one PRN chip) from the true code phase $\tau$. There is high cost associated with reporting a false acquisition rather than acknowledging a failure to acquire. Because of this, it is a standard practice to define a threshold $\theta_T$, and to report an acquisition only if $A_T(\hat{\tau}, \hat{f})$ exceeds $\theta_T$. The value of $\theta_T$ is determined by a specified probability of false alarm, which we define to be the probability that $A_T(\hat{\tau}, \hat{f})$ exceeds $\theta_T$ for at least one pair $(\hat{\tau}, \hat{f})$ for which $\hat{\tau}$ differs from $\tau$ by more than 1$\mu$s. Given a threshold $\theta_T$ and a value $\alpha^2$ of $C/N_0$, we define a probability of detection to be the probability that $A_T(\hat{\tau}, \hat{f})$ exceeds $\theta$ for some value of $\hat{\tau}$ within 1$\mu$s of $\tau$. We refer the reader to (Ward, 1996a; Ward, 1996b) for a more detailed account on probabilities of false alarm and detection and computation of acquisition thresholds.

Let us now discuss how the sensitivity of the first form of noncoherent processing (Equation (3)) compares with that of coherent processing (Equation (2)). We measure sensitivity here in terms of the lowest $C/N_0$, given any particular signal capture duration $T$, at which there is a threshold $\theta_T$ such that the probability of false alarm is $10^{-3}$ and the probability of detection is 0.5. Let us assume a Doppler range of $\pm 500$Hz, a code phase range of $[0, 50] \mu$s,

Figure 3 plots the performance attained for various durations of signal capture, for coherent processing and noncoherent processing with coherent durations $T/K$ of 9ms and 20ms. Observe that the sensitivity afforded by half a second of coherent processing is equal to that offered by two seconds of noncoherent combination of 9ms coherent blocks. Likewise,
one second of coherent processing offers the same level of sensitivity as twelve seconds of noncoherent combination of 9ms coherent blocks.

Clearly, Figure 3 points towards enormous sensitivity gains. But the performance advantages of coherent processing do not stop there – there are additional benefits to sensitivity and accuracy not captured by the simple model of signal acquisition defined above:

- In many urban environments, signals from satellites in one region of the sky are much stronger than others that are obstructed and therefore highly attenuated. If only the strong signals are acquired, the resulting ill-conditioned satellite geometry will lead to high dilution of precision (DOP) and degraded accuracy. In such situations, the sensitivity benefits of coherent processing result in a greater number of acquisitions and therefore improved DOP and accuracy.

- When received signal power levels vary significantly – as is the norm in urban environments – cross-correlations among satellite signals can lead to undesired peaks or distort a desired peak in the ambiguity function. This hurts sensitivity and accuracy of the code phase estimate. However, the effect diminishes with the length of coherent integration, becoming negligible when the duration of coherent integration exceeds a few hundred milliseconds. Hence, while cross-correlations detract from the performance of noncoherent processing, algorithms that process coherently can ignore them.

4.2 Computational Methods

We consider methods of coherent processing implemented on a digital processor. As such, we must work with a discretely sampled signal. We consider the model of a received signal introduced in Equation 1, except that we take $\gamma$ and $\omega$ to be band-pass filtered at $\pm 1.024\text{MHz}$.

\footnote{Conventional receivers track signals for tens of seconds, and over such long durations, differences in Doppler wash out the effect of cross-correlations on accuracy. This is not possible in A-GPS due to TTFF requirements.}
so that the signal can be represented by samples taken at the Nyquist rate:

$$x(\Delta k) = \alpha y(\Delta k - \tau)e^{-2\pi j(f\Delta k + \phi)} + w(\Delta k),$$

where $\Delta = 0.512\mu s$ and $k = 0, 1, 2, \ldots$.

The coherent ambiguity function is very closely approximated by

$$A_T(\hat{\tau}, \hat{f}) = \left| \sum_{k=0}^{[T/\Delta]-1} x(k\Delta)y(k\Delta - \hat{\tau})e^{2\pi jf k\Delta} \right|^2.$$

Signal acquisition algorithms typically compute and search over ambiguity values on a finite grid in the space of code phase and Doppler. To define a grid, let $\mathcal{T} \subset [\tau, \bar{\tau}]$ and $\mathcal{F} \subset [f, \bar{f}]$ be subsets consisting of code phases and Doppler frequencies that are multiples of “bin sizes” $\Delta_\tau$ and $\Delta_f$, respectively. We compute the coherent ambiguity value $A_T(\hat{\tau}, \hat{f})$ for each $\tau \in \mathcal{T}$ and $f \in \mathcal{F}$. An acquisition is associated with one of the ambiguity values on this finite grid exceeding a given threshold $\theta_T$.

There is a loss in sensitivity associated with discretization of the search – as $\Delta_\tau$ and $\Delta_f$ increase, sensitivity degrades. A reasonable choice of $\Delta_\tau$ and $\Delta_f$ are 0.512$\mu$s and $(1/4T)$Hz, respectively, which lead to a sensitivity loss of less than 1.4dB. (In the absence of noise, the ratio between the coherent ambiguity function at its peak and its value $0.256\mu$s and $(1/8T)$Hz away in the code phase and Doppler dimensions is approximately 1.4dB.) Note that, to maintain a reasonable degree of loss, $\Delta_f$ must be inversely proportional to $T$, and therefore the total number of grid points grows proportionately with $T$.

In the next subsection, we present some simple approaches to coherent processing and discuss why their computational requirements are prohibitive. There has been significant recent work geared towards reducing the computational requirements. We describe in Subsection 4.2.2 one algorithm representing some ideas explored over the past few years.

### 4.2.1 Simple Algorithms

Coherent processing algorithms have been studied for a long time in the GPS and broader spread-spectrum communications literature (see, e.g., Van Dierendonck et al., 1992; Van Dierendonck, 1996; Simon et al., 1994). The simplest approach to coherent processing involves use of time-domain correlation to compute ambiguity values. In particular, for each pair $\hat{\tau} \in \mathcal{T}$ and $\hat{f} \in \mathcal{F}$, this approach computes

$$\sum_{k=0}^{[T/\Delta]-1} x(k\Delta)y(k\Delta - \hat{\tau})e^{2\pi jf k\Delta},$$

and then squares. The compute time for this sum grows proportionally with $T$, and the number of such sums that need to be computed grows with $|\mathcal{F} \times \mathcal{T}|$, which is proportional to $T|\mathcal{T}|$. Hence, the total compute time is proportional to $T^2|\mathcal{T}|$.

As $T$ increases beyond a few milliseconds, time-domain correlation becomes too computationally demanding for a general embedded processor. One approach to accelerating
computations is to make use of an application-specific integrated circuit with many correlators that can compute the desired sums in parallel. An example of this is described in (Van Diggelen and Abraham, 2001). Here, the authors describe a circuit that offers about two thousand correlators per satellite. Hence, two thousand ambiguity values can be computed in parallel. If the ranges of code phase and Doppler uncertainties are 100μs and 500Hz wide, which are reasonable assumptions in some A-GPS architectures, and we use bin sizes of 0.512μs and (1/4T)Hz, respectively, then these correlators can accommodate T = 5ms of signal capture in real time. Hence, the strategy employed with this particular circuit is to carry out noncoherent combination of these short segments of coherently processed signal. One could try to extend such a strategy by increasing the number of parallel correlators in order to support longer durations of coherent processing, but this quickly becomes expensive and possibly infeasible for current integrated circuit technology.

Some alternative approaches to GPS signal acquisition make use of the fast Fourier transform (FFT) (see, e.g., Van Nee and Coenen, 1991; Moedlein and Krasner, 1998). Application-specific integrated circuits can also be designed to accelerate FFT computation. In fact, if designed right, an FFT circuit can compute a greater number of correlations in parallel than a circuit of the same size that performs time-domain correlation.

In one FFT-based algorithm, for each f ∈ F, the product x(kΔ)ej2πfkΔ is partitioned into 1ms blocks and these blocks are summed pointwise. The resulting vector of 2048 complex numbers is then correlated against the PRN via FFT-based convolution to obtain results for all 2048 code phases in parallel. The net compute time grows as T^2. Such an algorithm can offer significant benefits over time domain correlation. But as with time-domain correlation, compute time becomes unmanageable as T grows.

Another algorithm makes use of the FFT to compute ambiguity values across frequencies in parallel. This method is based on the observation that, given any code phase, the correlation values across frequencies can be viewed as a Fourier transform of x(kΔ)y(kΔ − τ), with appropriate zero-padding. Such an algorithm takes time that grows as |T|T log T, offering potential advantages over aforementioned algorithms as T grows large. Nevertheless, even this method poses onerous compute time requirements.

4.2.2 More Advanced Algorithms

In this section, we describe a more sophisticated algorithm for coherent processing, that captures some ideas explored in the recent GPS literature (Lin and Tsui, 1999; Lin and Tsui, 2000; Tsui, 2000; Psaki, 2001). This algorithm is not exact – it incurs a slight sensitivity loss relative to algorithms discussed in the previous section – but it requires orders of magnitude less computation for signal durations of interest.

For concreteness, let us assume the code phase to be within ±50μs and Doppler to be within ±200Hz. Let the duration T be a multiple of 1ms. Take Δ, and Δf to be 0.512μs and (1/4T)Hz. We partition the captured signal into blocks of 1ms duration, indexed by i = 0, 1, . . . , L − 1, where L = T/10^{-5} is the number of blocks. Let n = 2048 be the number
of samples per block. For each block, we define a correlation function

\[ z_i(\hat{\tau}) = \sum_{k=n}^{(i+1)n-1} x(k\Delta)g(k\Delta - \hat{\tau})\Delta. \]

We will approximate the coherent ambiguity function by

\[ \hat{A}_T(\hat{\tau}, \hat{f}) = \left| \sum_{i=0}^{L-1} z_i(\hat{\tau})e^{2\pi j \hat{f}ln\Delta} \right|^2. \]

This approximation incurs a sensitivity loss of less than 0.6dB. Let us describe an algorithm that computes this approximation:

1. For each \( i = 0, 1, \ldots, L - 1 \), compute \( z_i(\hat{\tau}) \) simultaneously for all \( \hat{\tau} \) using FFT-based convolution.
2. For each \( \hat{\tau} \in T \), zero-pad the vector \( (z_0(\hat{\tau}), \ldots, z_{L-1}(\hat{\tau})) \) to a length of \( 4L \), and then compute its FFT to obtain \( \sum_{i=0}^{L-1} z_i(\hat{\tau})e^{2\pi j \hat{f}ln\Delta} \), for \( \hat{f} \in \mathcal{F} \).
3. Square results from the previous step to obtain ambiguity values \( \hat{A}_T(\hat{\tau}, \hat{f}) \) for \( \tau \in T \) and \( \hat{f} \in \mathcal{F} \).

Asymptotically, this algorithm takes time that grows proportionately with \( |T| T \log T \). However, this term is scaled by a small constant, and in the range of interest - up to a few seconds - the dominant term influencing compute time grows linearly with \( T \). This leads to enormous reductions in computational requirements, as illustrated in Figure 4. This figure compares the approximate number of compute cycles used by the various algorithms we have discussed. The unit of measurement corresponds to the amount of time it takes to execute one complex-valued multiplication and one complex-valued addition. It is assumed that a \( K \)-point complex-valued FFT takes \( 4K \log K \) compute cycles. In each case, the compute time estimate is associated solely with acquisition of a single satellite signal.
It is interesting that – despite having been a long-standing problem – the state-of-the-art for coherent processing of GPS signals left much room for innovation. The advances we have discussed in this section represent significant strides in technology for coherent processing. However, they still do not measure up to contemporary challenges, which call for acquisition of possibly ten or more satellite signals through coherent processing over durations of more than a second, all within seconds of compute time. Proprietary research conducted during the development of UrbanGPS™ has lead to further advances that address these demands.

4.3 Modeling and Further Computational Challenges

The signal model we have presented is a simplification of reality. Real GPS signals are distorted by various phenomena, including satellite and Earth acceleration, receiver dynamics, receiver oscillator instabilities, and multipath and fading effects. Many of the complexities arising from such phenomena need not be taken into account when designing algorithms that coherently process short durations of signal and combine results noncoherently. However, as the duration of coherent processing increases, algorithms become sensitive to these effects and break down. Novel signal models are required to accommodate coherent processing over seconds of GPS signal. Such models give rise to a number of additional computational challenges beyond those posed by coherent processing with a simple signal model, and efficient algorithms are required to accommodate the complexities of these models. Much research effort has been geared towards design of appropriate signal models and companion algorithms.

5 Information Fusion

In A-GPS, information from multiple sources can be used to enhance satellite signal acquisition. The receiver starts with constraints that limit the uncertainty about variables of interest, including time, receiver location, velocity, and acceleration, and oscillator characteristics. Furthermore, as satellite signals are acquired, each acquisition provides additional information about these variables. In this section, we discuss a computational approach for fusing available information to enhance sensitivity.

5.1 Nonlinear Filtering

As information is gathered and processed, it can be optimally fused to aid further acquisition through nonlinear filtering. The idea is to maintain and update a probability distribution over eight variables of interest: three-dimensional receiver location, time, three-dimensional receiver velocity, and receiver time velocity, by which we mean the rate of change of receiver time with respect to a standard time source. We will refer to this eight-dimensional space as the geometric search space. The probability distribution represents our beliefs at any stage during signal processing. As more information is processed, the probability distribution is adapted according to Bayes’ rule. This approach to information fusion has been explored extensively in the engineering literature (see, e.g., Elliott et al., 1994) and has also
been proposed in the context of GPS signal acquisition (Spilker, 1996b, p. 298). However, implementation has been impractical due to onerous computational requirements. In the next section, we describe geometric search – a efficient approach to information fusion that approximates the performance of nonlinear filtering.

5.2 Geometric Search

Geometric search operates on the geometric search space defined in the previous section. At any time in the processing, geometric search represents uncertainty in terms of polyhedron in this space. This polyhedron defines a region of feasible points – points outside this region have been ruled out by information gathered up to the current time.

Initially, the polyhedron is generated by linear constraints deduced from assistance data. For example, upper and lower bounds on time offered by the time stamp lead to linear constraints. The base station coverage area can also be approximated in terms of linear constraints. Bounds on Doppler can be linearized to form additional linear constraints.

As signal processing ensues, whenever a satellite signal is acquired, additional constraints are introduced. In particular, an acquisition constrains the intervals of feasible code phases and Dopplers for a satellite, and this translates into linear constraints on the geometric search space. Furthermore, as more data is processed, the Doppler uncertainty associated with an acquired satellite signal shrinks, tightening the associated constraints on the polyhedron.

Figure 5 illustrates feasible sets maintained by geometric search. The map on the far left depicts a two-dimensional projection of an initial feasible region. Upon the acquisition of a first satellite signal, new information is incorporated and the feasible region shrinks, as portrayed by the middle map. As more satellites are acquired, the region continues to shrink, ultimately leading to a small feasible region, as shown in the map to the far right. (Note that the map on the far right is magnified relative to the others.)

At intervals of time during processing, the polyhedron is used to generate constraints on code phase and Doppler. Each constraint is efficiently generated by solving a linear program (see, e.g., Bertsimas and Tsitsiklis, 1997). For example, an upper bound on code phase is
produced by maximizing the code phase – which is a linear function of the geometric search space – subject to the polyhedral constraints on the geometric search space. Such constraints enhance sensitivity since the acquisition threshold associated with a given probability of false alarm decreases with the uncertainty in code phase and Doppler.

6 Anti-Multipath Triangulation

In urban environments, multipath is the dominant factor contributing to accuracy degradation. This section reviews traditional methods for multipath mitigation, explains why they do not adequately accommodate urban environments, and describes a new approach, which we call anti-multipath triangulation.

6.1 Conventional Approaches to Multipath Mitigation

After computing pseudoranges, conventional GPS methods typically employ weighted-least-squares triangulation to estimate receiver location. This implicitly assumes that pseudorange error distributions are Gaussian. With this technique, multipath-induced biases in pseudorange estimates can lead to large location errors. Traditional approaches to multipath mitigation focus on detecting multipath-induced biases and alleviating their influence on the triangulation process. Two broad approaches explored in the literature are:

- **Receiver Autonomous Integrity Monitoring (RAIM).** RAIM assesses consistency among pseudorange estimates and discards those that are significantly inconsistent from the majority. We refer the reader to (Parkinson and Axelrad, 1988; Brown, 1992; Walter and Enge, 1995) for more detailed discussions of RAIM.

- **Peak separation.** These methods typically aim at identifying and using the code phase associated with earliest signal to be received, when there are multiple correlation peaks. We refer the reader to (Van Dierendonck et al., 1992; Garin et al., 1996; Garin and Rousseau, 1997; Calh and Chansarkar, 1997; Veitsel et al., 1998, McGraw and Braasch, 1999) for more on this topic.

Conventional methods of multipath mitigation are quite effective in environments with open-sky exposure. For example, in aviation – the primary driving application for these methods – it is common to acquire a direct signal from each of eight to twelve overhead satellites, and it is reasonable to model multipath as a single reflection off of a single object. Such a model is well served by RAIM and/or peak separation methods.

In the process of developing UrbanGPS\textsuperscript{TM}, we have collected and analyzed enormous quantities of GPS data from major cities around the world, including Tokyo, Seoul, New York, and San Francisco. Through this statistical study, it has become clear that urban environments pose a form of multipath entirely different from that accounted for by conventional models. It is common to receive only reflected signals and the delays introduced by these reflections can be severe. Traditional methods like RAIM and peak separation often fail to improve accuracy in such situations.
Figure 6: Anti-multipath triangulation reduces errors by about a factor of two relative to conventional weighted-least-squares triangulation. The 68.3 percentile error is 104m with conventional triangulation and 55m with anti-multipath triangulation.

6.2 A New Model and Algorithm

An alternative approach to dealing with multipath in urban environments involves development of a statistical model that accurately captures behavior of pseudorange error distributions observed in real data, such as asymmetry, fat-tails, interdependence among errors, and dependence on multiple observable factors. Given this statistical model, location can be estimated by maximizing likelihood or minimizing expected squared error. If error distributions were Gaussian, this would be equivalent to solving a weighted-least-squares problem. However, complexities associated with an accurate statistical model make estimation a computational challenge. Anti-multipath triangulation is a proprietary technology developed to address this challenge. This algorithm solves the maximum likelihood problem in less than ten milliseconds on a typical server or a few hundred milliseconds on an embedded processor.

Through cross-validation tests, anti-multipath triangulation has proven effective in real urban environments with severe multipath. In Figure 6, results based on weighted-least-squares triangulation are compared against anti-multipath triangulation. Each scatter plot represents location fixes over a series of independent trials at the same location. The receiver was on the 29th floor of a 50-story office building in Tokyo, about 13 meters from the closest window. With weighted-least-squares triangulation, multipath leads to poor accuracy – errors are on the order of a hundred meters. Anti-multipath triangulation reduces error by about a factor of two.

7 Time Stamp Recovery

Our discussion of A-GPS has revolved around methods that make use of a time stamp within hundreds of microseconds of the true time. When such an accurate time stamp is not available, time stamp recovery algorithms are required. Syrjärinne (2000) provides a nice overview of approaches to time stamp recovery, which he partitions into three classes:
1. **Base station synchronization.** A network operator can synchronize base stations with GPS time by incorporating location measurement units (LMUs), which are essentially GPS receivers that interface with base stations.

2. **Navigation data alignment.** In A-GPS, navigation data is received as assistance information. By aligning this data with the signal received from a satellite, transmission time can be determined, and an accurate time stamp can be deduced from this.

3. **Enhanced triangulation.** It is easy to formulate a problem of generating location and a time correction from a coarse initial time stamp and either code phases or navigation bit transition times. Such a problem can be solved by more sophisticated versions of triangulation, which we refer to as enhanced triangulation.

   Syrjärinne (2000) points towards enhanced triangulation as the approach of choice. Significant work has been directed towards development of associated computational methods (Camp, 1997; Syrjärinne, 2000; Syrjärinne, 2001; Sirola and Syrjärinne, 2002; Akopian and Syrjärinne, 2002). Though this work represents important progress, there are some drawbacks to the general approach of enhanced triangulation:

   1. Enhanced triangulation methods that have appeared in the literature require acquisition of five satellite signals for robust operation (Camp, 1997; Syrjärinne, 2000; Syrjärinne, 2001; Sirola and Syrjärinne, 2002; Akopian and Syrjärinne, 2002).

   2. Coherent processing algorithms require an accurate time stamp to operate efficiently. In the enhanced triangulation approach, acquisition is carried out prior to time stamp recovery. This precludes coherent processing over long durations.

   3. When pseudorange errors are large (e.g., a hundred meters are more), as is common in urban environments, enhanced triangulation methods can lead to erroneous estimates of satellite positions. This amplifies the error introduced by urban multipath.

   Despite shortcomings, Syrjärinne (2000) finds enhanced triangulation to be preferable because of obstacles presented by alternative approaches. Base station synchronization is impractical because deployment of LMUs requires a prohibitively large investment. As for navigation data alignment, Syrjärinne (2000) identifies the following as obstacles:

   1. Alignment of navigation data requires a correlation process that can lead to false acquisitions and therefore erroneous time estimates.

   2. Navigation data alignment presents computational challenges.

   3. Such an approach requires an initial time stamp within a few seconds of the true time, because only a few seconds of navigation data are available to the receiver.

   4. Navigation data alignment is not reliable when $C/N_0$ is low.

   We argue that these obstacles can be overcome and that navigation data alignment is a very effective approach to time stamp recovery. The first obstacle disappears if one sets the correlation threshold to appropriately bound the probability of a false alarm. There are sophisticated algorithms that can address the second issue. The third hurdle can be overcome by making greater amounts of navigation data available to the receiver and using the fact that much of the navigation message repeats each 30 seconds. In the event that sufficient
navigation data can not be made available, one might employ enhanced triangulation to recover a better time stamp, and then follow up by applying navigation data alignment to compute a more accurate time stamp and estimates of satellite and receiver locations.

As indicated by the fourth item, sensitivity suffers when the time stamp is inaccurate. More specifically, efficient navigation data alignment calls for $C/N_0$ to be at a level that enables reliable acquisition with noncoherent summing of coherent blocks of about 10ms in duration. Given two seconds of signal capture, the $C/N_0$ required is about 6dB higher than that which would be required for signal acquisition with coherent processing and an accurate time stamp. Fortunately, this bears little noticeable impact on receiver sensitivity as observed by the end user, for reasons that we will now explain.

It is important to note that the $C/N_0$ requirement for navigation data alignment is satisfied if any one of the received satellite signals is sufficiently strong. This is because time stamp recovery can be accomplished using a single satellite signal. Once this time stamp is recovered, the remaining satellite signals can be processed coherently. As illustrated by the histograms of Figure 1, in challenging urban and indoor environments there is typically significant variance among $C/N_0$ offered by different satellite signals. A location fix generally requires that four satellite signals offer $C/N_0$ sufficient for acquisition by coherent processing algorithms. It turns out that in most real-world scenarios where this happens, the strongest satellite signal is at least 6dB stronger than the weakest one acquired. Hence, the use of navigation data alignment to recover a CDMA-grade time stamp does not significantly affect the ability of the receiver to acquire a location fix.

References


