

Efficient accurate positioning Why radios need ToA capability

September 2014 (updated 2018)
David Bartlett



Introduction

Radio signals can be used to determine the positions of objects relative to one another. However, not all measurements made by different radios and wireless devices and systems are equal, and not all measurements lead to useful positioning performance.

This paper provides a tutorial style review of what is needed in measurements in order to make them useful for the purposes of positioning and locating things. It reviews signal strength (RSSI) and Time-of-Flight (ToF) techniques, but the emphasis is on Time-of-Arrival (ToA) because this leads to the best performance: better accuracy, faster update rates, better scalability, and lower power consumption.

The objective is to provide an understanding of radio measurements and what is needed to compute a position fix using Time-of-Arrival processing techniques.

Background

In WP001 we introduced the concept of relative positioning using collaborative techniques as used in the Series 500 Cluster location system.

In a Series 500 wireless mesh sensor network, devices (nodes) broadcast radio signals (referred to as chirps) to neighbours within radio range. By measuring the precise Time of Arrival (ToA) of chirps received a node is able, with the help of the data included in the chirp, to compute its position relative to its neighbours. The Series 500 system is based on a CSS radio PHY (IEEE 802.15.4a).

WP002 describes how the Omnisense Joint Timing and Location Engine (JTLE) works assuming that sufficient measurements with adequate quality can be obtained from the radio.

Definition

In this paper we refer to ToA (time of arrival). Readers will undoubtedly be familiar with industry terms such as TDOA (time difference of arrival); OTDOA (observed time difference of arrival), E-OTD (enhanced observed time difference), all of which refer to time differences rather than simply Time-of-Arrival.

Various techniques for differencing (and double differencing) times are used by the LE (Location



Engine) in the process of computing positions. However as far as the radio is concerned: It has a local clock, usually unsynchronised, which it uses to measure the time of arrival of a radio signal from a neighbour. In its most basic form this is what we call the Time of Arrival measurement. Any differencing carried out after making the measurement is algorithmic and not core to the subject of this paper.

The critical thing at the radio level is to be able to measure the time of a signal (arrival of a received signal and sometimes transmission of another signal) with sufficient precision for the measurement to be useful. It is also a requirement that the clock used is sufficiently stable over the measurement interval of interest to ensure that clock errors don't degrade performance.

Therefore in this paper we are concerned with techniques for measuring the time of arrival of a radio signal using a local free-running unsynchronised clock and for this reason we use the simple term Time-of-Arrival. There are two parts:

- How to measure signal time-of-arrival;
- Requirements for clock stability and management.

Principles for ToA to work

In a ToA system all (or some) of the devices (nodes) transmit a radio signal, which we refer to as a chirp or signal, which any authorised neighbour within

radio range can receive, measure the time of arrival and decode the payload. Nodes take it in turn to transmit - this may be managed within a TDM (time division multiplex) structure, or use an “Aloha” (random access) approach, or other suitable scheduling algorithm.

This is a one-to-many broadcast architecture. The messages may be part of the protocols for data flow within the network and don't need to be different or specific to the localisation task.

There are a few general sufficiency requirements in order to be able to compute reliable positions:

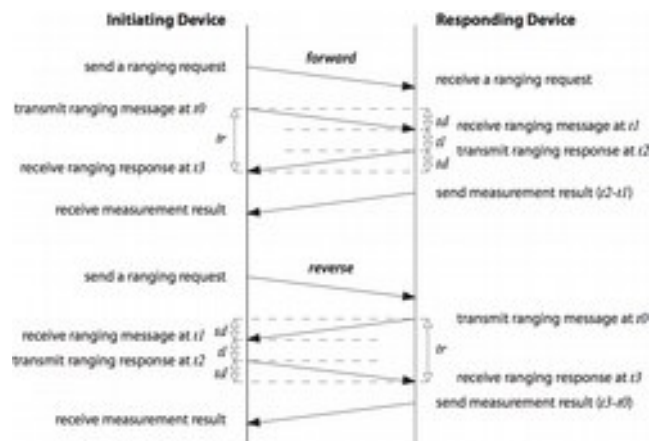
- Devices need to be within range of sufficient neighbours - in general measurements of four or more are required for a 3D position fix (3 for a 2D fix), but depending on the architecture more may be required. It is a good principle to be able to receive at least twice the minimum number required;
- It is helpful to the LE if devices can both transmit and receive, but this is not an essential requirement;
- Accuracy with which Time of Arrival is measured is critical. This is determined by the bandwidth of the signal, S/N ratio, integration time, measurement method and fidelity of the measuring circuitry;
- A sufficiently stable clock is required;
- Transmit timing may be as important as measuring Time of Arrival, but it is generally easier to implement.

A Time-of-Flight scheme

A pair of radio devices can measure the distance separating them using time-of-flight (ToF). This involves an exchange of messages between them in which the round-trip time is measured and the turn around time subtracted to yield the double range measurement.

The figure illustrates how a ToF range is typically measured. It usually involves both forward and reverse measurements which are averaged to eliminate the effect of relative clock drift between the two devices.

The ranges measured between neighbours are input to the LE which is able to compute their relative positions. Since the relative clock offsets and drifts were removed at the measurement stage the LE only needs to solve for positions and not clock offset parameters. Therefore, compared with ToA, fewer neighbour measurements are required. Nevertheless the simple ToF approach scales poorly with network size: it scales $O(N^2)$, whereas the broadcast ToA method scales $O(N)$. Therefore ToF methods are slower and require more power than ToA. They are



often less accurate as well, because there is less scope for the LE to optimise the solution based on the measurements obtained. ToF is better suited to infrequent on-demand positioning than continuous (X,Y,Z) positioning.

Using Signal Strength (RSSI)

Most radios and radio chips are able to provide a Received Signal Strength Indicator measurement (RSSI). Since signal strength reduces the further the signal travels, signal strength represents an estimate of distance for positioning.

Signal strength varies non-linearly with distance, so RSSI is a better indicator of range when devices are close to one another. Most current Wi-Fi and non-GPS cellular positioning solutions as well as RFID and iBeacon systems are based on RSSI, very often combining measurements from several transmitters with signal strength maps (or finger prints) which characterise the expected signal strength across the area of interest.

RSSI-based techniques are seldom accurate, being more suited to proximity determination than continuous (X,Y,Z) positioning. Signal strength is strongly affected by many factors outside of the users' control:

- signal obscuration, such as by one's body, and other changes to the environment, can introduce order of magnitude changes to signal strength;
- multi-path causes both constructive and destructive interference leading to measurements that can be much larger or smaller than expected;
- antenna radiation patterns and polarisation are not isotropic so the orientation of the receiver relative to the transmitter can introduce very large signal strength changes
- transmitted power levels vary and receiver measurement accuracy is usually quite low.

Measuring ToA

The Time-of-Arrival of a radio signal is typically measured in one of two different ways:

- A broadband signal with a known envelope is cross correlated with a signal template by the receiver. The peak of the cross correlation output represents the time offset between the two signals, the observed Time-of-Arrival. This method is often used with spread spectrum signals (CDMA), including GPS, or other wideband signals such as CSS (chirp spread spectrum) or UWB (Ultra Wideband). It is may also be useful to know something about the time of transmission, sometimes called ToD (time of departure).
- By measuring the phase of a signal or carrier, the measured phase represents the observed Time-of-Arrival. In this case the time of arrival is actually the delay relative to the start of the phase cycle and the measurement repeats every wavelength, leading to the so-called cycle ambiguity.

These two methods have different characteristics in the presence of multipath and NLoS (non-line-of-sight) signal propagation.

Cross Correlation for ToA

For a broadband radio signals the usual method for measuring the time-of-arrival is to construct a signal template representing the known transmitted signal and to cross-correlate this template with the actual received signal.

In signal processing, cross-correlation is a measure of similarity of two waveforms as a function of a time offset applied to one of them. It is also known as a sliding dot product or sliding inner-product. It is commonly used for searching a long signal for a shorter, known feature. For the purposes of measuring ToA, the longer signal is the (continuously) received signal and the shorter known feature is the template of (a portion of) the transmitted signal.

Therefore ToA is usually measured on the known defined portion of the transmitted signal: for example the pre-amble or training sequence or the header portion of transmitted signal. For the rest of the signal which is unknown (variable data content) it is much harder, and sometimes impossible, to accurately and reliably measure ToA.

The accuracy with which one can determine the peak of the cross-correlation function depends on a number of factors:

- the bandwidth of the signal;
- time interval over which the cross-

correlation is performed;

- signal-to-noise ratio of the received signal;
- frequency offset;
- channel impulse response.

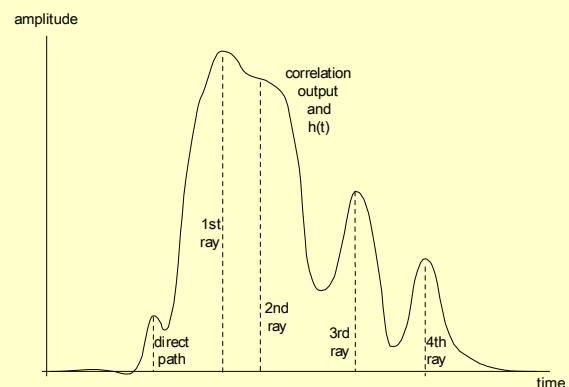
The greater the bandwidth of the signal the higher the resolution in time. There is an inverse relationship between bandwidth and time. The Shannon-Nyquist theorem describes this principle. In simplified terms it states that a signal of bandwidth B needs to be sampled at a minimum of $2B$ samples per second in order to encode all the information in the signal. As a corollary this means that sampling a band-limited signal at more than $2B$ samples per second does not contribute any additional information. Therefore the fundamental time measure for a band-limited signal is $1/2B$. This means that, by way of example, a 10 MHz bandwidth signal has a basic sampling time resolution of 50 ns which given the speed of propagation of radio signals at the speed of light represents a distance travelled of 15 metres.

There are ways of increasing the apparent resolution by interpolating the output of the cross-correlator, recognising that the peak falls between samples. The extent to which this interpolation succeeds depends not only on an understanding of the shape of the peak, but also for how long one integrates and what the S/N ratio of the received signal is.

It is also very dependent on the nature of the radio environment. Usually the signal arrives via multiple paths: reflected and refracted, and one ends up cross-correlating with a complex distorted

Multipath and ToA

Cross correlating the received signal against a signal template gives the channel impulse response, an example of which is illustrated below. Multipath and other environmental effects lead to a complicated result and the challenge is how to determine which is the earliest arriving version of the signal, not necessarily the biggest, because it is most likely to be the direct path.



representation of the original signal. See boxout “Multipath and ToA” for a simple illustration of this effect. Of course what one wants to find is the signal that has travelled by the direct path (earliest arrival) and for many indoor environments it is simply not present, or not strong enough to measure, or it is masked by other close-in-time signal paths. A lot of research effort has gone into better algorithms for measuring the channel model from which a better estimate of the actual earliest ToA can be obtained, but in practice one is usually dependent on the silicon in the radio receiver chip for making these measurements and in order to keep power requirements and complexity (cost) down many compromises are made.

Phase measurement

For signals that have a continuous phase envelope, such as the signal carrier, or pilot tones modulated onto the carrier, phase measurement can be used as an alternative to a broadband cross-correlation. Phase measurements can give extremely high resolution (and precise positioning), but they are subject to the cyclic ambiguity of the signal wavelength and are also easily degraded by multipath.

Under normal operating conditions: reasonable S/N ratio, limited multipath, and cost-effective signal sampling, it is generally possible to measure the wavelength of a signal to around 1% of the wavelength (rule-of-thumb). So if we have a carrier of 1 GHz, this means a wavelength of 3 m and a measurement resolution of 3 cm. Therefore we should be able to position to an accuracy of a few centimetres, but for any distance greater than 3 m the number of wavelengths the signal has travelled is ambiguous and may limit the practical usefulness.

Many practical systems using phase measuring employ multiple wavelengths acting together like a vernier. Note that the signal whose wavelength is being measured does not need to be the signal carrier, it can be a modulation of the carrier, or, as in the Omnisense SWB system, obtained from the differences between multiple carriers.

How accurately can we measure?

The best theoretically achievable performance of positioning systems is estimated using the Crámer-Rao Lower Bound (CRLB) which is a statistical metric defining the best measurement possible using an unbiased estimator of the variable. See boxout for description.

In addition to bandwidth, time and S/N there are also a number of practical considerations too:

- The stability of the clock;
- The fidelity of the ToA measuring circuits

and algorithms;

- Other radio measurements, specifically frequency offset and Doppler;
- Number of neighbours used in the position computation, and the geometry of the solution;
- Extent to which multiple signals over time-space can be integrated;
- Capability within the radio system to detect measurement errors;
- Environmental factors such as multipath and interference - often the most telling factors of all;
- Phase and time ambiguities.

Crámer Rao Lower Bound

Named after the mathematicians Harald Crámer and Callyampudi Radhakrishna Rao, the Crámer Rao Lower Bound (CR-LB) is a technique from the field of Estimation Theory based on Maximum Likelihood that defines the lower limit to variance that can be attained by an unbiased estimator of a parameter of a statistical distribution.

Even though measurements used to compute position in positioning systems are seldom Gaussian and often not linear either, the CR-LB has become a common method for attempting to calculate the best theoretical performance that can be obtained for a positioning system.

Given system performance specifications the CR-LB has been derived for several common positioning techniques:

Wideband system based on cross-correlation:

$$\sigma \propto \frac{1}{\sqrt{\left(\frac{S}{N}\right)BT}}$$

where B is the bandwidth and T the integration time and (S/N) the signal to noise ratio.

For phase measuring systems:

$$\sigma \propto \frac{1}{2\pi f_c \sqrt{2\left(\frac{S}{N}\right)T}}$$

In practice most systems do not achieve performance close to the CR-LB, largely because of other system non-linearities and environmental effects such as multipath and NLoS.

Clock Stability and management

For ToA based positioning solutions to be reliable it is imperative that the clocks at both transmitter and receiver are properly managed and sufficiently stable.

Suppose a device transmits a signal received by four neighbours at known positions which measure the ToA of this signal. Provided the clocks on the receivers are synchronised the position of the transmitter can be solved: four measurements used to solve the four unknowns of X,Y,Z and T (time), where T is the unknown clock offset of the transmitting device.

Whilst theoretically possible to synchronise the clocks of the devices at known positions, and many of today's commercial system do this (including GPS although satellites are transmitters not receivers), it is costly and complex and advantage can be gained from systems which do not require physical synchronisation of devices. The easiest way of achieving the same result is to pseudo-synchronise devices using one or more LMUs (local measurement unit). An LMU is simply a transmitter (or receiver) at a known position. Therefore using the ToA measurements it is possible to calculate the clock offsets of the devices at known positions using the LMU signal.

However, the LMU does not transmit at the same time as the mobile unit, and the clocks drift during the intervening period. Provided the clocks can be adequately modelled and tracked this doesn't matter. For most real-world practical systems the period of predictability of the clock is important. Referred to as the clock "coherence" time, it is typically around 1 to a few seconds for low cost commercial TCXOs. This means that in 1 second the clock will remain within a small enough tolerance of that predicted by the clock model. Typically we are looking for one or at most a few nanoseconds of unpredictable drift per second of elapsed time. (3 ns means approximately 1 m measurement error)

The other requirement on the clock is the ability to stop or start it when required. In order for two measurements made at different times to be used in a position calculation the clock must run uninterrupted over the time interval spanning the first and second measurements. However, in order to save power we wish to shut down the clock when it is not needed. The Omnisense JTLE is able to cope with a clock that is interrupted between groups of measurements.

Transmit Timing

For a single isolated signal transmitted as described above it does not matter what the transmit time is. However, if we chose to transmit a sequence of signals close together in time (within the clock

coherence period) or if the devices are also able to receive and measure the arrival times of signals from neighbours then the LE can use this information to integrate coherently and produce significantly better results, and knowledge of transmit time becomes important. The Omnisense Series 500 system works in this way.

Knowing the transmit time does not necessarily mean measuring it in the same way as for received signals, it is usually easier to arrange that signals are transmitted synchronously with the start (or a known part) of the local clock. The time base for the local clock is usually chosen to be short, but long enough to avoid range ambiguity. Frequencies in the range of Hz to MHz are commonly used.

For the best possible positioning performance it is essential that multiple measurements can be made and coherently combined.

GPS has a basic code repetition interval of 1 ms, which is assembled into data bits and frames, so even though the basic correlation code repeats every 1 ms each data bit is 20 ms and measurements can be placed at a unique time on a much longer time scale. This level of sophistication is not needed for most local positioning systems.

However the choice of base time interval is a fundamental parameter in the design of the radio.

Generalised ToA Approach

For ToA based positioning to work we need to satisfy a number of conditions according to the specific requirements of the application:

- A reference clock with an unambiguous interval long enough to satisfy the maximum ranges used: given radio propagation at the speed of light this equates to around 300 m per microsecond.
- Each device in the network has its own version of this clock – they do not need to be synchronised, but the stability of each clock must meet the required coherence time: every 3 ns of unpredictable clock drift introduces approximately a metre of error.
- Ability to leave the clock running continuously through groups of measurements or to shut it down to save power.
- A signal with measurable characteristics such as a known signal envelope for correlation or a defined phase behaviour that is transmitted periodically.
- Signal transmissions should be at a known time: either by measuring time of transmission, or, more usually, by arranging for the transmission to be triggered by a defined clock event such as the beginning of

the clock interval for the reference frame.

- A receiver that can accurately measure the times of arrival of signals from neighbours: 3 ns of measurement error translates into about a metre of position error.
- The ability to measure auxiliary radio parameters: in particular RSSI and frequency offset (Doppler) are useful.

The last two points are actually the hardest to do, but the first five pertaining to clock management and signal transmission are often overlooked.

Radio requirements

To build a viable high-accuracy local positioning system we need a radio (chip) that can deliver the following:

- Clock with suitable stability. Low cost commercial TCXOs can achieve coherence times of 1 second or more at the 1 metre level.
- Suitable measurement time base: of the order of 10 μ s (3 km) is adequate for many local and indoor positioning systems.
- The ability to shut the clock down or allow it to run continuously through several transmit and receive measurements.
- Signal transmission aligned to the base clock boundary, or a defined sub-multiple of it, or for the time-of-departure to be measured and transmitted with the radio signal.
- A suitable training sequence or defined header that can be used as a template for measuring ToA.
- The ability to transmit a data payload in all messages, including those for which ToA is measured.
- The ability to measure accurate ToA's, to an accuracy of at least one tenth the inverse bandwidth, preferably better.
- Alternatively, or in addition, the ability to measure carrier phase, or the phase of signal sub-carriers or pilot tones.
- Ability to measure RSSI and frequency offset (Doppler) are highly desirable.
- All measurements made relative to a single free-running clock (same for transmit and receive).

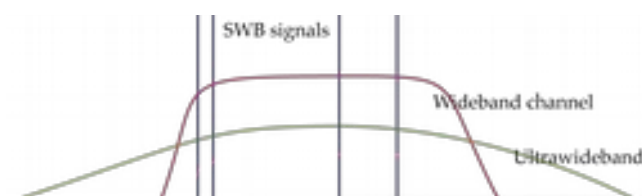
The Omnisense Location Engine can and does make use of all measurements; in particular it builds on sequences of signals transmitted and/or received with a continuously running clock, including devices that are able to both transmit and receive, as well as non-radio measurements from inertial sensors to

yield superior positioning performance.

Case Study: Sparse Wide Band

Omnisense has coined the term SWB (sparse wide band) for a phase measuring technique that uses the measured phases of multiple sub-carriers within a wideband channel. Whilst the first generation¹ radio was proprietary and used hopping between narrow band channels, the technique is one of the approaches that can be used for measuring ToA in OFDM (orthogonal frequency division multiplex) systems as used in many modern communications networks, as well as one that could be applied to narrow-band radio systems.

SWB uses phase differences between pairs of sub-carriers within a multichannel system (generated from the same base clock). By using multiple pairs of carriers with different frequency spacing one has different wavelengths for which the phase is being measured. The longest wavelength is defined by the base time interval and it defines the maximum distance that can be measured unambiguously. Larger frequency differences are used to gain higher precision. This approach is simple and highly effective. It solves the cycle ambiguity problem, gives excellent accuracy and also eliminates unknown phase offsets that are traditionally introduced into the signal carrier by the mixer and up-conversion LO in the radio transmitter.



The core parameters of the first generation SWB radio are:

- 50.78125 kHz time reference and channels;
- Nearly 400 usable sub-carriers in a 20 MHz wide channel (2.4 GHz ISM band);
- Approximately 1 ms measurement time;
- Four sub-carriers giving differences of approximately 700 kHz, 2 MHz and 10 MHz;
- Chirp repetition rates of up to 10 Hz;
- Simple Aloha contention scheme;
- Hopping between sub-carriers to minimise cost and complexity of tag;
- Frequency offset (Doppler), RSSI and signal quality measured and reported.

1 The first generation SWB radio was developed for a particular application (sport) and is not in general production.

The location engine used close carrier spacings for coarse positioning and the largest spacings were used to compute the final high resolution position. Multiple chirps from each tag are combined to give robust time-space diversity to the solution, with performance at the 0.5 m level outdoors, over ranges up to 1 km. Doppler and signal quality were used to provide positioning confidence and a direct estimate of velocity.

This SWB implementation is equally applicable in OFDM systems in which phase measurements of the pilot sub-carriers are used in the JTLE instead of the hopped tones of the bespoke SWB radio described above. Most standards-based OFDM radio PHY's have sufficient pilot channels and the channel spacing between pilots tends to be well suited for positioning use.

Another of the major strengths of the system described is the fact that it is frequency agnostic. Even though the first implementation uses 2.4 GHz, the same approach could be used in any suitable frequency band, subject to the usual constraints of bandwidth and radio regulations.

Examples of likely performance

Ultrawideband

The best performance is likely to be achieved with UWB systems. As standardised in IEEE 802.15.4f the channel bandwidth is 500 MHz which gives good multipath resolution for use indoors, although higher frequencies and low powers limit the ability to penetrate solid objects and restrict optimum performance to line-of-sight (or near LoS) situations. For these systems performance at the 0.1 m level is achievable.

Wi-Fi and wide channel systems

Wi-Fi (IEEE 802.11a) and IEEE 802.15.4a using wideband channels of 80 MHz or 160 MHz should be able to achieve reliable 1 m performance, although the results will be somewhat degraded in significant multi-path situations: probably only realistically able to achieve around 3 metres.

Operating bands are 2.4 GHz or 5 GHz and although greater ranges than UWB can be achieved they will be limited indoors by poor penetration through solid objects. The Omnisense Series 500 product is based on IEEE 802.15.4a.

Broadband sub-GHz systems

For the best penetration through solid objects, while preserving enough signal bandwidth the sub-GHz whitespace spectrum provides a potential solution. Here we have channel bandwidths of 8 MHz (6 MHz in the USA). With careful radio design it is possible to achieve performance at the 1 m level

outdoors, although in heavy multipath situations accuracy is expected to degrade more noticeably to, perhaps, around 10 m.

Narrow-band systems, 2.4GHz

IEEE 802.15.4 defines a PHY with around 3 MHz of bandwidth used primarily for low-rate telemetry systems, often using protocols such as 6LoPAN, ZigBee, ISA100 and others. It should be possible to achieve performance at the 3 m level outdoors, although indoors in heavy multipath environments this is likely to degrade significantly.

Narrowband sub-GHz systems

For narrow-band systems performance is likely to be even lower, and it is probable that they will have to rely on phase measurements rather than signal cross-correlation. These systems often have channel bandwidths of 25 kHz to 200 kHz. With such narrow channels realistic direct ToA measurements are unlikely to be possible. However, by measuring carrier phase, particularly if it can be arranged to combine multiple channels at different frequencies, positioning at a 10 m level should be possible.


Conclusions

The future of ubiquitous positioning for IoT and consumer applications needs to progress beyond basic RSSI proximity measurements. ToF is a solution for small systems and one-off position calculations but for real scale the use of ToA will be required.

Industry has steered away from ToA, largely because it is complex and difficult to do well, but the Omnisense JTLE is changing this. All that is needed now is a richer eco-system of radios and radio chips capable of measuring and reporting ToA measurements.

About Omnisense

Omnisense Limited is a Cambridge UK based IoT business. Based around an advanced Location Engine Omnisense solves the tough problem of locating and positioning things and making information available for developers to build location aware applications.


Omnisense Limited
Unit 1B, Alington Road
St Neots
Cambridgeshire
PE19 6WL
UK
+44 (0) 1223 911 197
info@omnisense.co.uk
<http://www.omnisense.co.uk/>
knowhere anywhere