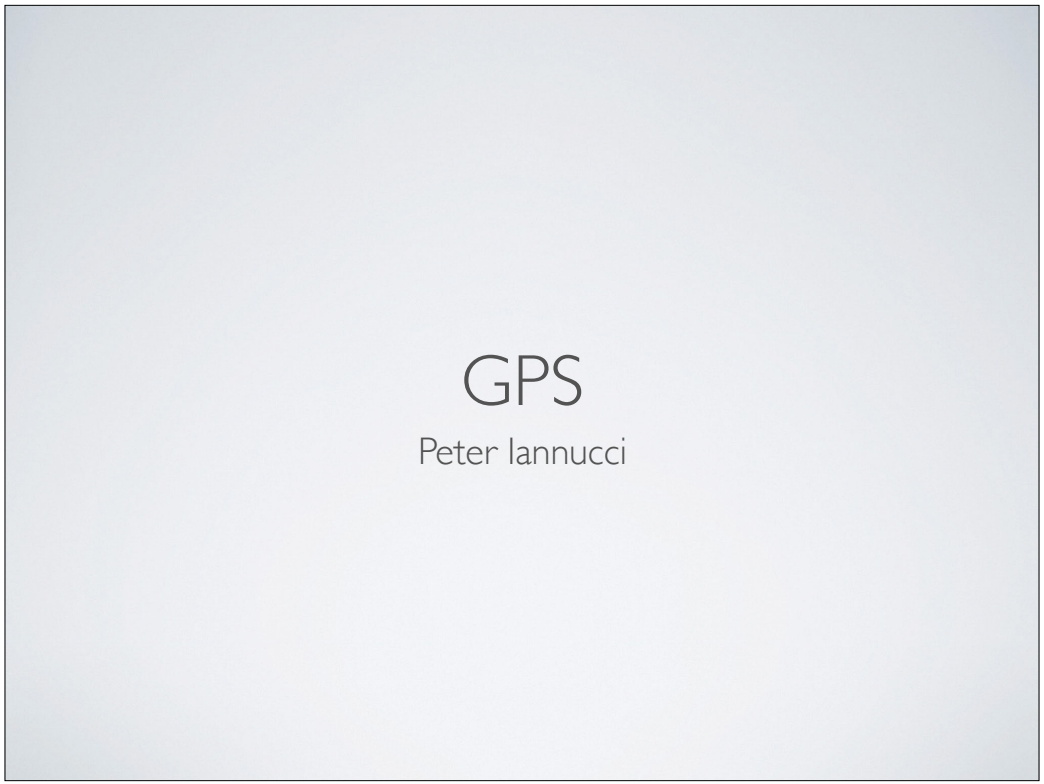
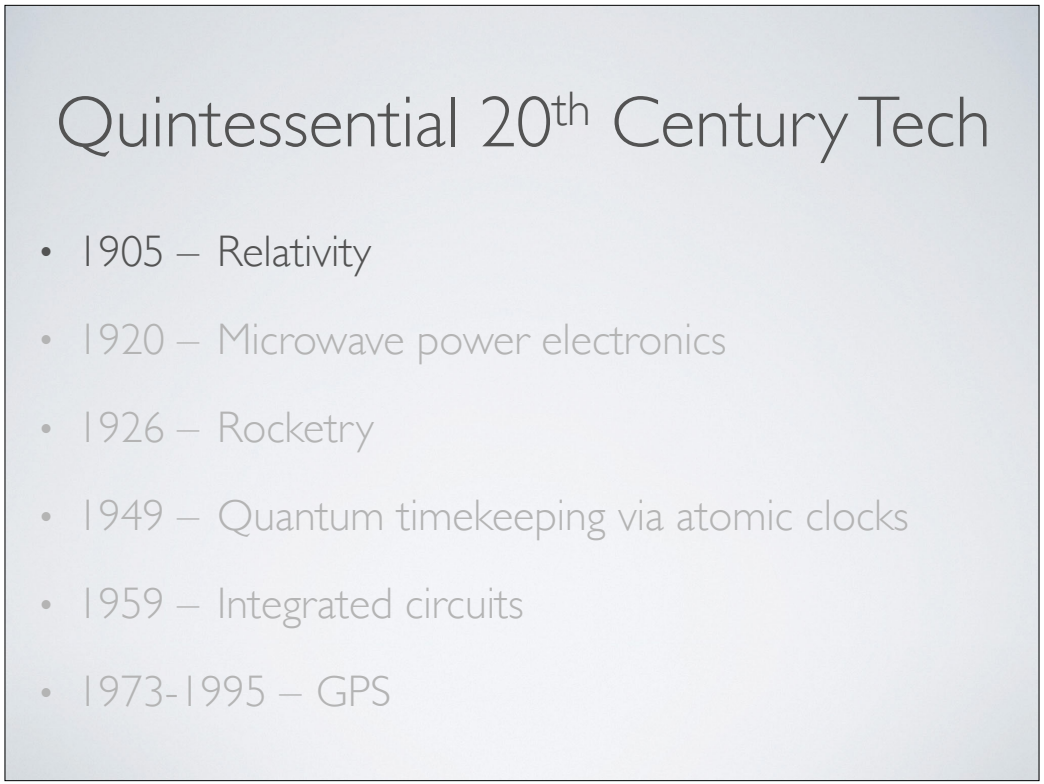


Good morning! I'm excited to talk with you about the Global Positioning System. I'll give some background on the origins and functionality of the system, and then we'll talk about how it works, and how that impacts mobile applications.

We use the term GPS, but a Global Positioning System is really a type of system; there's more than one such system in existence. The specific implementation that we use every day is called NAVSTAR, or Navigation System Using Timing and Ranging. NAVSTAR is a project of the US Department of Defense, and it's a synthesis of a number of earlier systems.

A slightly different term that you'll hear more and more in the future is GNSS, or Global Navigation Satellite System. GPS has come to refer specifically to the American system, so GNSS is the preferred internationalized term going forward.

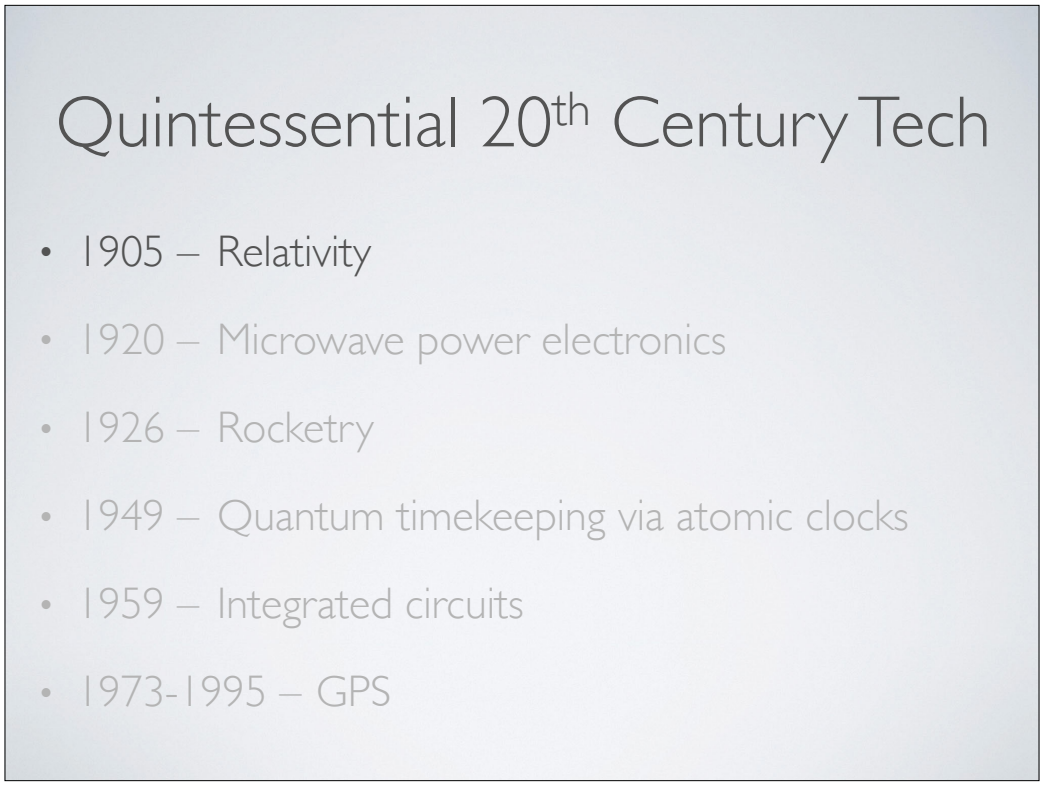


GPS is the quintessential 20th century technology. It integrates developments from physics and electrical engineering in a really remarkable way. In the last lecture, Sam talked a bit about the science of Earth measurement, or geodesy. A big problem in geodesy is establishing a consistent set of coordinates in which to record all types of geographic data. The coordinate system for GPS comes from Einstein's theories of special and general relativity, from 1905 and 1915.

Ideally, we would set up coordinates so that points on the surface of the Earth don't move. Having the surface of the Earth stay pretty much in one place is an important assumption you have in mind when you sit down to draw a map.

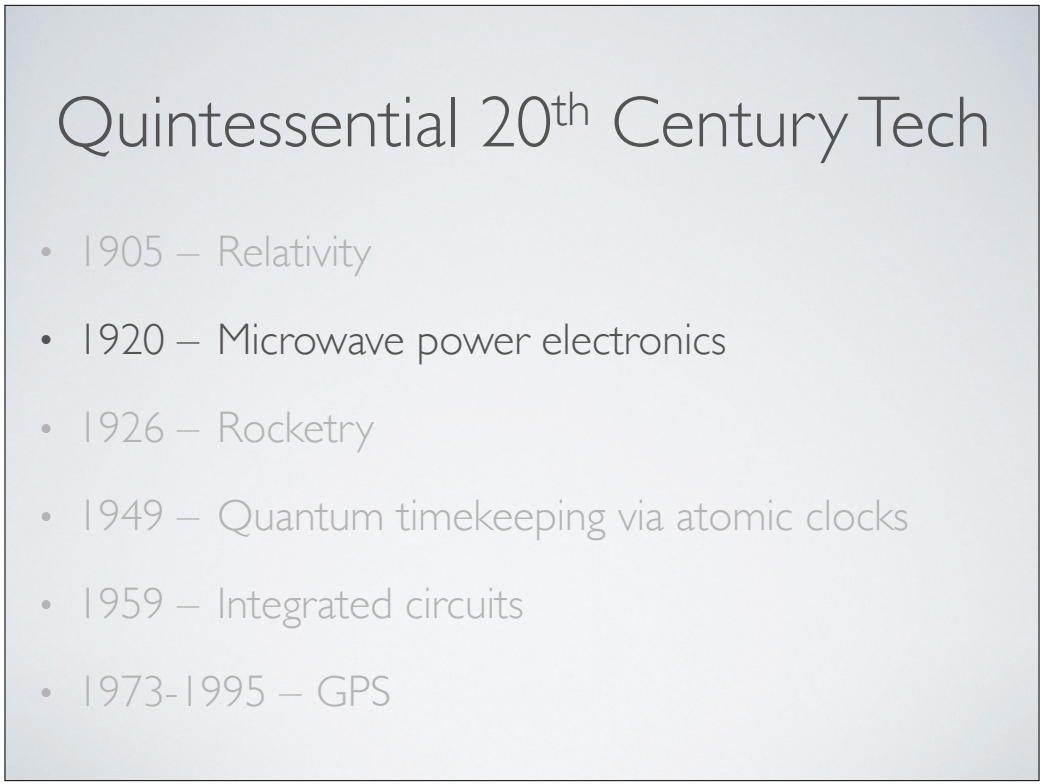
<draw> We can do that; we can put the origin of our coordinates at the Earth's center-of-mass, and we can have the z axis sticking out through the north pole, and the x axis sticking out through the equator at the prime meridian, and so on. In this frame of reference, points on the Earth have fixed coordinates over time, so the frame is called Earth-Centered, Earth Fixed.

Ultimately, though, we are going to be making measurements with clocks, not rulers. Measuring distances with clocks is much better because rulers the size of the planet don't fit in your pocket. So while Earth-Centered, Earth Fixed coordinates are perfectly good coordinates for measuring distances, we also need a coordinate system for time.



It turns out, though, that in general relativity, it's not possible to define a consistent notion of time in a spinning frame of reference. If you synchronize two clocks on the Earth, and carry one all the way around the equator, the clocks will disagree by 200 nanoseconds. If uncorrected, this would lead to distance errors of about 200 feet. To make matters worse, due to the Earth's gravity, at higher altitudes, clock run *slightly* faster than at lower altitudes.

So GPS calculations are performed in Earth-Centered Inertial coordinates, which are fixed with respect to the distant stars, and GPS devices measure time in the Earth-Centered Inertial frame. The GPS satellites have all had their clocks corrected for the difference between their altitude and the surface of the Earth.



Another development that enabled GPS was the development of high-powered microwave transmitters in the period between the two World Wars. GPS satellites transmit about 45 Watts of power each. Compared to your cellphone or your Wi-Fi, that's quite a bit of power, but considering how far away these satellites are, it's remarkable that we can pick up these signals at all. If it weren't for the development of microwave amplifier devices in the 20's and 30's, it wouldn't be possible for the satellites to generate signals strong enough to pick up.

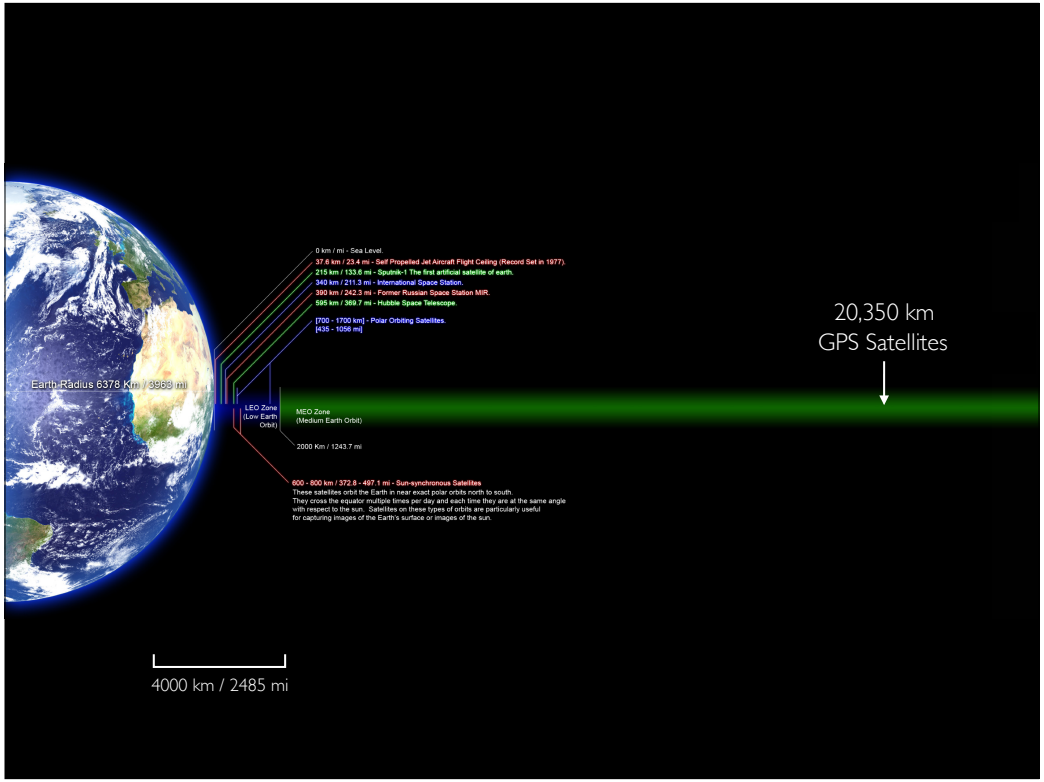
Just to give you an idea of how small this 45 Watts number is, that means the entire collection of 30 satellites only transmits as much power as a single toaster oven. Their energy is spread out over the entire surface of the Earth and beyond, so only a tiny fraction is available to my mobile device at any given moment. My phone has to pick up a signal of less than 10^-15 watts.

That means the receiver circuitry has to pick out a signal that's only 200 nanovolts, compared to 1300 nanovolts for the random thermal fluctuations of the electrons in the antenna. This is hard to do, and requires a lot of processing. That's one reason why turning on the GPS receiver in your phone will drain your battery.

Quintessential 20th Century Tech

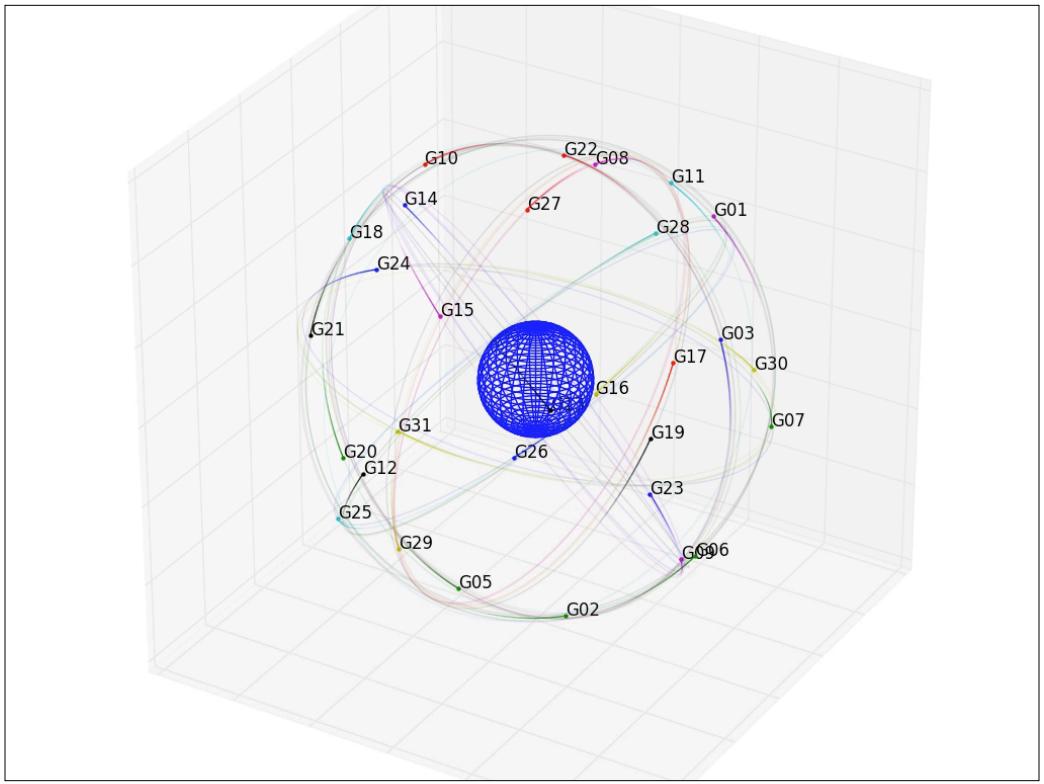
- 1905 – Relativity
- 1920 – Microwave power electronics
- 1926 – Rocketry
- 1949 – Quantum timekeeping via atomic clocks
- 1959 – Integrated circuits
- 1973-1995 – GPS

These satellites are in medium-earth orbit, far from the atmosphere and the gravitational disturbances of large landmasses. They orbit the Earth twice per day.

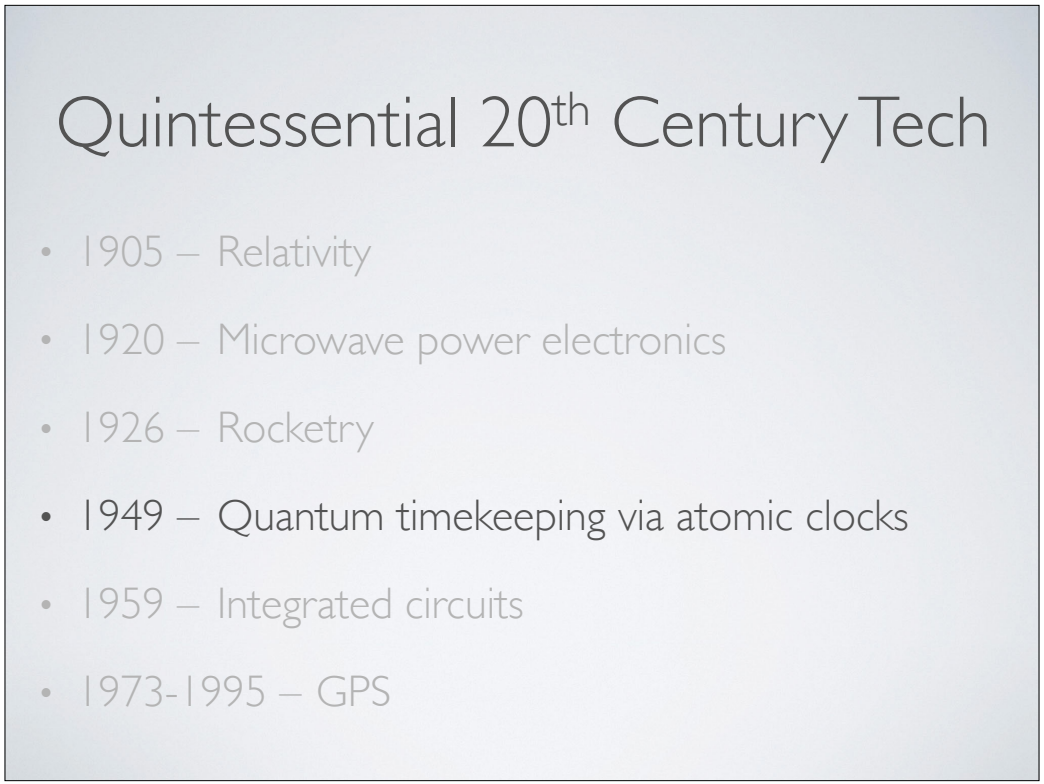


The GPS satellites fly at about seven Earth radii, far above the International Space Station and where the space shuttles used to fly. They're about 2/3 of the way to geosynchronous orbit.

Placing the satellites at such a high altitude makes them visible from an entire hemisphere at a time <draw it>. This requires fewer satellites to provide coverage to the entire planet than would be needed if they orbited at a lower altitude. It also synchronizes the satellites' orbital periods with that of the Earth, so that after two orbits each satellite returns to the same place in the sky as viewed from any location. I've also prepared an animation so that you can see how the entire constellation of GPS satellites is laid out.

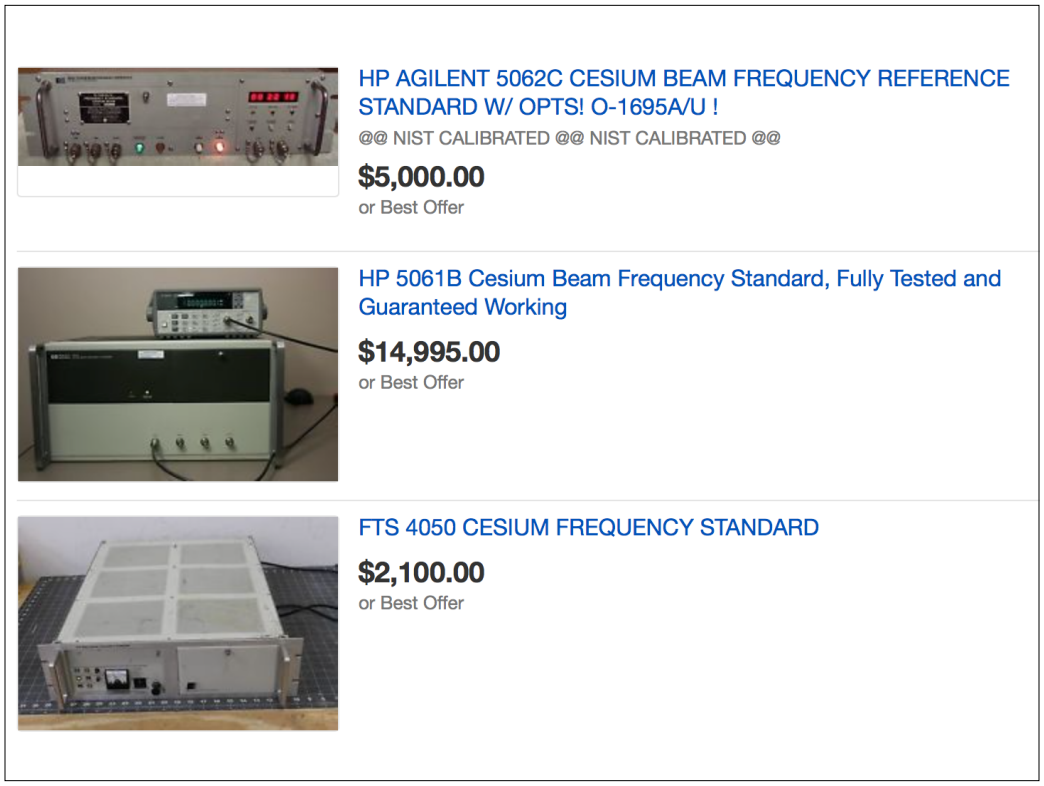


Here are the locations of the satellites right now, based on data I downloaded from NASA an hour ago. The satellites are organized into six groups, called planes, which each cross the Earth's equator at a different point. This arrangement is a compromise that provides good coverage in most parts of the world, and ensures that the satellites pass over lots of US-allied countries where we can build control stations. The blue sphere shows the size of the planet Earth for scale. Each satellite carries a numeric designation, which we'll see later is used to distinguish their signals.



Atomic clocks are another ingredient that made GPS possible. These devices use the quantum mechanical behavior of cold atoms to construct clock oscillators that keep extremely faithful time. We're talking better than 10 parts per trillion accuracy over time scales of a few seconds, and half a part per trillion over time scales of a few hours.

Each GPS satellite has four atomic clocks of two different types, which provide both redundancy and robustness. Atomic clocks use a lot of power, typically 40-50 watts, and the GPS satellites collect several hundred watts of solar power to keep the clocks running.



Atomic clocks are also bulky and expensive, which is why it makes sense to put the fancy clock in the satellite rather than in the mobile phone. That said, it was actually somewhat fiddly to put an atomic clock into orbit without shaking it to pieces on the launch pad, and to make it so that the device didn't require any service. Modern GPS satellites have design lifespans of 15 years, but the original model of satellite was only designed to last for 5 years.

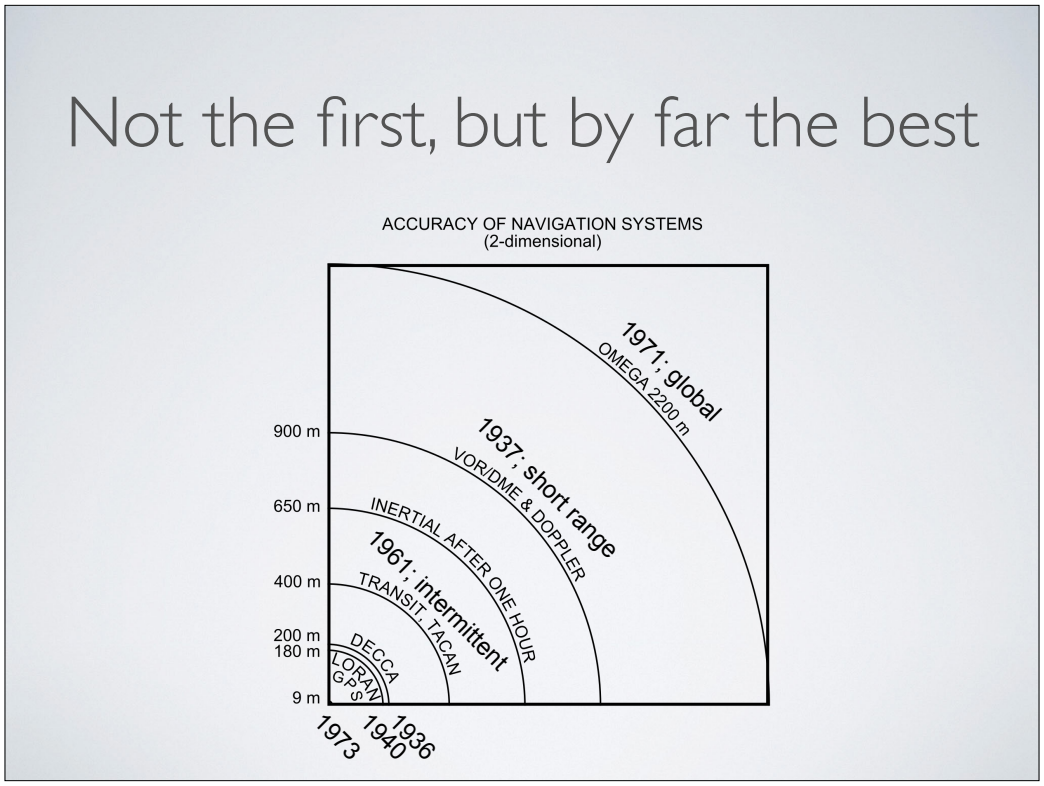
- ~ 120mW power consumption
- ~ 17cm^3 volume
- 35g weight

As an aside, in the past few years, chip-scale atomic clocks have become available. The folks who build these devices have been singing about how these are going to revolutionize GPS accuracy, and I'm sure they're right, but for now this thing is 20% the size of my entire phone, one quarter the weight, and uses quite a bit of power. It doesn't help that they cost $1,500.

Quintessential 20th Century Tech

- 1905 – Relativity
- 1920 – Microwave power electronics
- 1926 – Rocketry
- 1949 – Quantum timekeeping via atomic clocks
- 1959 – Integrated circuits
- 1973-1995 – GPS

Let's get back to the big picture. GPS was made possible by a number of earlier developments. It was also preceded by a number of other radio navigation systems.



I won't say much about these, except that none of these earlier systems provided localization services as precise, fast, or widely available as GPS. With a few exceptions, these systems have all been discontinued over the past 20 years in favor of GPS.

There's one more thing I want to discuss before I tell you how GPS works. In order to put the design into perspective, you need to understand what GPS was designed to do.

Function

1. *Zoom+boom: Directing nukes, aircraft, ground forces*
   - Global coverage (guaranteed up to 3,000 km altitude; partial up to 36,000 km altitude)
   - Selective availability (disabled, Clinton, 2000)
   - Anti-spoofing (active; limits civilian uses)
   - GPS a military system operated by US Air Force
   - Error target is 95% ≤ 2.6-11.8 m (global average)
   - Designed to outlast its operators… by a few months

GPS was originally called the Defense Navigation Satellite System, and its purpose was to help deliver American nuclear missiles and bombs to their targets. It was conceived of as a force multiplier, which is to say that one on-target nuke is as effective as many off-target nukes.

As a consequence, GPS has global coverage — not just of populated regions, but also of regions where intercontinental ballistic missiles might fly, like 3,000 kilometers above the north pole.

GPS was also designed with controllable errors and signal distortions built-in, so that the United States could deny the use of this system to its adversaries. This was called selective availability, and it introduced errors of about 100 meters to civilian GPS locations until it was switched off by order of President Clinton. The Department of Defense has promised never to turn it back on, and they've announced that the next generation of satellites won't even have the capability any more.

In order to prevent an adversary from generating a fake GPS signal and confusing our nukes, GPS was also designed with an anti-spoofing technique. Legitimate military GPS signals are encrypted with a certain code that our adversaries are supposed to be unable to predict. Special military GPS receivers can be programmed to use this encrypted signal in addition to or in place of the unencrypted civilian signal, but everybody else is stuck with the less-precise, spoof-able, unencrypted signal.

Function

1. *Zoom+boom: Directing nukes, aircraft, ground forces*
   - Global coverage (guaranteed up to 3,000 km altitude; partial up to 36,000 km altitude)
   - Selective availability (disabled, Clinton, 2000)
   - Anti-spoofing (active; limits civilian uses)
   - GPS a military system operated by US Air Force
   - Error target is 95% ≤ 2.6-11.8 m (global average)
   - Designed to outlast its operators… by a few months

The original satellites were pretty dumb. Since they were launched in 1978, and were designed to operate in a harsh radiation environment for five years, their logic was exceedingly simple and robust, to the point where the core signal generation logic could be implemented with perhaps a few hundred gates. All the data processing necessary to track the satellites' locations and monitor the health of their on-board clocks was done on the ground. So the signal transmitted by the satellites, called the navigation message, which carries information about where the satellites are and what corrections the receiver should apply, is uploaded from the ground stations every 10 hours or so and repeated verbatim by the satellites.

That's why, if you want the best accuracy, you should look at the GPS signal immediately after the ground stations upload fresh navigational data, while the data is still fresh. After a few hours, unpredictable effects begin to accumulate and the clock and orbit information becomes increasingly inaccurate. So, ignoring atmospheric effects, the military GPS signal is expected to be accurate to within 2.6 meters all over the world 95% of the time immediately after a fresh upload, and within 11.8 meters just before the next upload.

This accumulation of errors would continue if the ground stations went offline, to 380 meters after two weeks, and many kilometers after 180 days. The unclassified performance document says, "Such a condition could occur as the result of total loss of the Control Segment due to a natural or man-made disaster". Food for thought.

Function

2. *Directing civilian aviation*
   - Prevent disasters like Korean Air Lines Flight 007 (shot down in USSR airspace; no survivors)
   - 95% ≤ 6.0-12.8 m (global average) for civilians
3. *Directing the pizza delivery guy*
   - Great strides made in civilian receiver accuracy

Civilian navigation — the thing we care about — is a secondary function of GPS, added by the order of President Reagan. This decision was made in the aftermath of the destruction of Korean Air Lines Flight 007 in 1983. The plane strayed into Soviet airspace due to a navigation error, and 269 lives were lost.

The civilian navigation signal of GPS has somewhat lower accuracy than the military signal, with error targets of six to 13 meters, depending on the time since the last data upload.

Satellitenpolitik

- GNSS is a weapon; US anxious to maintain control
- Rocket launch systems = ICBMs; US seeks to limit tech transfer
- GPS is a target for jamming, if not anti-satellite attacks, in the event of war
- Alternatives created by other countries
  - USSR (GLONASS), EU (Galileo), China (Běidǒu/COMPASS)
  - On-going spats about transmission frequencies, modulations

The United States has been jealous of its control over this technology, and it has not been terribly encouraging to other nations seeking to deploy their own global navigation satellite systems.

But those nations have gone ahead anyway, and some of the competing systems are quite interesting in their own right. GLONASS is the most operational of the group, and thanks to a Russian tax on imported GPS devices unless they support GLONASS too, it's actually pretty widely supported. This is neat because it means the GLONASS constellation can be used in addition to the GPS constellation, meaning that we get 60 satellites total. Instead of having maybe five satellites overhead at any given time, we get 10. The Galileo and Běidǒu constellations will, hopefully, be joining the party soon.

How does it work?

- PRN codes – encoding time into the signal
- 4-lateration – computing positions from times
- Acquisition and tracking – finding weak signals
- Almanac – finding satellites quickly

I hope I haven't bored you with all of this background information. This is a technology class, so let's talk tech. How does GPS actually work? There are four pieces you need to understand.

First is how the signal from the satellite encodes the time it was sent.

Second is how the receiver can use the times that signals were sent from four different satellites to learn where it is and what time it is.

Third is how the receiver can find these incredibly weak signals that are smaller than the noise.

And fourth is how the receiver can be spared the cost of searching constantly for weak signals.

Encoding time into the signal

- Coarse/Acquisition (C/A) code — civilian
  - Changes on microsecond timescales
- Precise (P(Y)) code — military
  - Changes on 100 ns timescales

The satellite signals come in two flavors. The first is the Coarse/Acquisition code, or C/A code, and the second is the Precise code, or P code. The P code is almost always encrypted, and of limited use to civilians. But you should know that it exists.

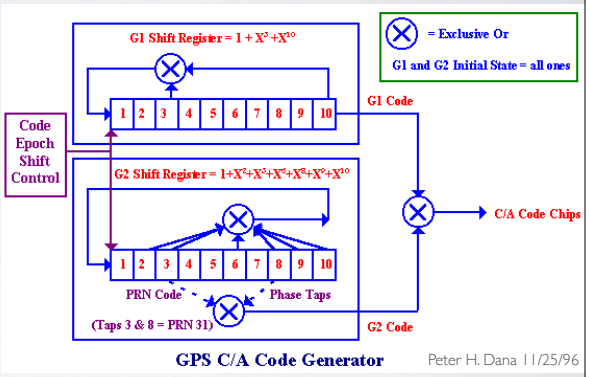
Encoding time into the signal

- $\text{C/A Signal} = \text{Carrier} \cdot (-1)^{\text{chipping sequence} + \text{nav. message}}$
- **Carrier** = sinusoid at e.g. 1,575.42 MHz
- **Chipping sequence** = 1.023 Mbits/s, pseudo-random 1023-bit sequence repeats every 1 ms
- **Nav. message** = 50 bps, data
  Clock corrections, orbit information, satellite health

The civilian C/A code is transmitted as a 1.5 GHz sinusoid with a series of rapid phase reversals, where the phase reversals are due to two factors. The first is a chipping sequence. This is just a repeating, pseudo-random pattern with some special properties that we'll talk about in a minute. The second factor is the navigation message. This carries a digital signal that repeats every 12.5 minutes, telling the receiver how to correct for clock errors, where the satellite will be at any given time, and whether the satellite has any fault conditions that the receiver should know about.

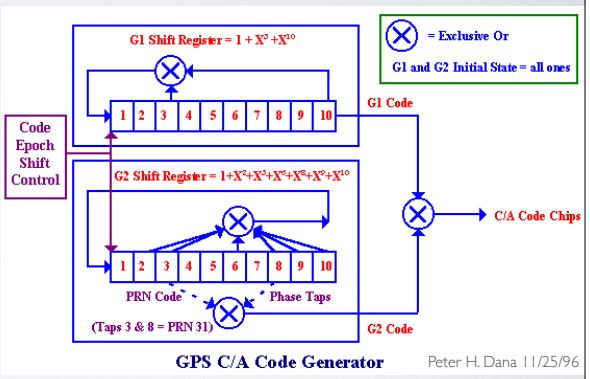
PRN codes

- Linear Feedback Shift Register (LFSR)
- Bits move to the right, one step per cycle
- First bit assigned XOR of other bits
- All bits initially 1
- Polynomial representation — contact with Galois theory

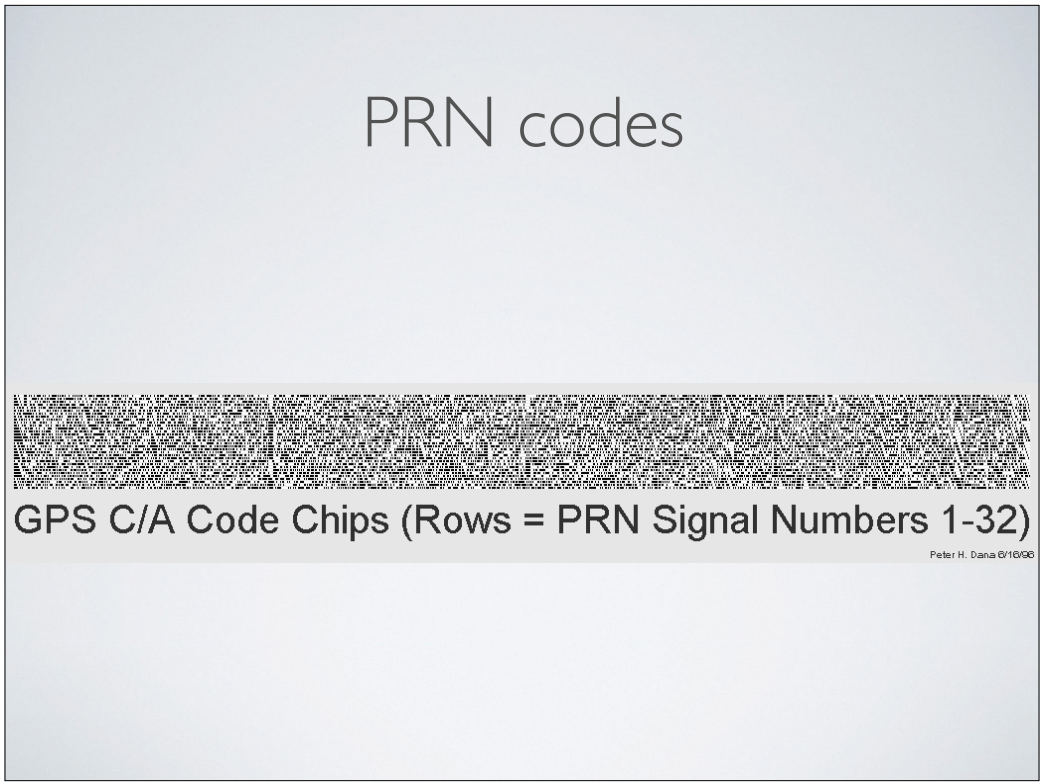


PRN codes

- 2^n-1 non-zero states form a cycle
- Zero state leads to itself
- Choices of bits to XOR for output of G2 code
  - Family of "Gold codes"
  - One per satellite



The key property that led the designers of GPS to choose these codes is that it is difficult to mistake any two codes for each other, and it is difficult to mistake one code for a shifted copy of itself. This means we can "recognize" the pattern of a particular satellite at a particular phase, and immediately know the transmission time (modulo 1 millisecond). We can then use the navigation message to tell which millisecond is which, and obtain the absolute transmission time.



Here's the result of running that circuit forward 1023 times, until it repeats. Each row corresponds to a different choice of the bits to XOR, and hence a different satellite.

4-Lateration

- Our four unknown coordinates are $(\vec{\mathrm{x}},\mathrm{t})$
- Signal i announces its origin from $(\vec{\mathrm{x}}_i,\mathrm{t}_i)$
- Signals travel at speed of light: $|\vec{\mathrm{x}}-\vec{\mathrm{x}}_i|=\mathrm{c}(\mathrm{t}-\mathrm{t}_i)$
- Each visible satellite gives one equation

Okay, now we have a way to encode the transmission time into the satellite signal, via these repeating patterns at the sub-millisecond timescale and the navigation message at longer timescales. Now we can pose the central GNSS positioning problem: given the observed satellite signal timestamps, and given the navigation data from the satellites telling us where they are in their orbits, how do we determine where we are?

4-Lateration

- Four equations suffice to determine (x,y,z,t)
- Non-linear, non-convex, over-constrained: nasty
- Define RMS $\mathbf{error}(\vec{x},t)$ as
  $\sum_{\text{satellites}} |(\text{speed of light}) \cdot (\text{signal time-of-flight}) - (\text{estimated distance})|^2$
- Solve by iteration
  - Guess $(\vec{x},t)=(0,0)$
  - Compute slope of error with respect to unknowns
  - Reduce error using Newton's method, i.e. $x_{i+1} \coloneqq x_i - f(x_i)/f'(x_i)$

Sources of Error

- Weak or reflected signals
- Inconvenient geometry
- Ionospheric distortion
- Transmitter clock errors – can be deliberate
- Transmitter location errors – can be deliberate

The GPS signal is already weak, so a small amount of absorption along the path between the user and the satellite is a big problem. Signals following an indirect path from the satellite also produce misleading data. For instance, the signal could bounce off the ground, or off a building, before arriving at the receiver. The problems of absorption and reflection are particularly bad in urban canyons — the spaces between tall, downtown buildings.

One thing that helps is that the signal is designed to have circular polarization. That means the fluctuating electric and magnetic field vectors run around in a circle, rather than simply oscillating back and forth through zero. When a circularly polarized signal reflects off the ground or a building, it changes from right-handed to left-handed and vice versa. A good GPS receiver will be designed with an antenna that only accepts right-handed signals — but antennas like that are usually not small enough to fit in a mobile device.

A separate problem is inconvenient satellite geometry. <draw> If your satellites all happen to be near each other in the sky, then large changes in your location result in small changes in your observed range data. So your uncertainty in your range measurements is amplified, and your uncertainty of your position is large. Conversely, if your satellites are well-spaced in the sky, then your measurements are as sensitive as possible to your position, and your precision is not diluted.

Sources of Error

- Weak or reflected signals
- Inconvenient geometry
- Ionospheric distortion
- Transmitter clock errors – can be deliberate
- Transmitter location errors – can be deliberate

Receivers have to account for the fact that GPS signals are distorted and delayed as they pass through the ozone layer and the rest of the ionosphere. Otherwise, your position estimates will be off by many meters. <pause> The ionosphere is a region of the atmosphere where ultraviolet light from the sun ionizes air molecules. These ions absorb part of the electric field of the GPS signal and radiate their own, slightly delayed signal. <pause>

It's possible to model these errors pretty well if we have dual-band measurements, because the effect of the ions depends on the frequency of the signal. This is why GPS transmits on two frequencies, called L1 and L2. For now, civilians can only get this information in a limited way, since the L2 transmission only carries the encrypted military signal. As an alternative, local ionospheric data can be obtained through other channels, for instance through the Internet.

I mentioned before that the Department of Defense put in a feature they called Selective Availability to deny the use of GPS to our adversaries. The way they implemented that was by introducing deliberate errors into the clock and location information provided by the satellites.

Sources of Error

- Weak or reflected signals
- Inconvenient geometry
- Ionospheric distortion
- Transmitter clock errors – can be deliberate
- Transmitter location errors – can be deliberate

Even with Selective Availability turned off, the navigation data provided by the ground stations and retransmitted by the satellites becomes inaccurate over time. Not every effect that changes the trajectory and behavior of the satellites can be predicted or modeled, so it's hard to know in advance exactly where the satellites will be. Surveying applications can take advantage of retroactive data that pins down the locations of the satellites to within a few centimeters. This data is published online hours or days after the fact, after they collect and merge observations from around the world. That's where I got the data for the animation I showed you earlier.

Acquisition and Tracking

- Correlation
- Doppler
- Compute-heavy
- Much easier after initial acquisition
  - Only look in adjacent correlation "bins"

Okay; we know how the satellite signal is structured to carry time information, and we know how to use time information to find where we are. But how do we pick up the signal if it's much weaker than the noise?

The way we're going to solve this problem is by integrating, or averaging, the energy from the satellite over a period of time. Say we average together *n* individual measurements. The standard deviation of this average will go down like the square root of *n*, because the variances of independent variables add, like this: (chalkboard)

This is great, because it means that if we average for a long enough period of time, even a very weak signal will be revealed as the noise dies out. But remember that the GPS signal is constantly flipping its phase by plus-or-minus 1. If we're not careful, all those pluses and minuses will wash out in our sum, leaving nothing. So the acquisition computation will have to exactly match the phase of the satellite signal, or else the signal will disappear completely into the noise.

Acquisition and Tracking

- Correlation
- Doppler
- Compute-heavy
- Much easier after initial acquisition
  - Only look in adjacent correlation "bins"

By the way; because we are comparing the received signal against our expectations of when the phase transitions will occur, our straightforward averaging procedure is referred to as a correlation technique.

This requirement that we need to exactly predict the phase transitions of the satellite signal is both a blessing and a curse. It is a blessing, because it means that when we find the signal, it will show up very strongly at one particular phase and not at all at other phases. This will allow us to estimate the transmit time very accurately. But it is also a curse, because it means that we have to compute 1000-plus different averages before we can hope that one of them will be significantly nonzero.

In fact, the situation is worse than this, because the satellites are moving so quickly that their signals are substantially Doppler-shifted. This means that the receiver will have to compute 40,000 different averages, each using upwards of 1,000 samples. This calculation can be accelerated using the Fourier transform, but it's still expensive. This is why starting up the GPS on your phone takes time and costs a lot of power.

Almanac

- Satellite ephemeris (plural: ephemerides)
- Warm/cold start
- 12.5 minutes for a full download

Orbit parameters allow receiver to estimate Doppler shifts and to guess which satellites will be visible, saving a great deal of effort.

Additional Topics

- Assisted GPS
- Differential GPS
- Relative Kinematic Positioning
- Augmentation systems

Making a good thing better

- Regional augmentation systems (satellite- or ground-based)
  - WAAS (North America), EGNOS (EU), MSAS +QZSS(Japan), India, …