1

Historical Development

1.1 Introduction

If you board a commercial flight in 2016, you will step onto an aircraft that has a significant redundancy of electrical power and safety systems with a high level of automation. The instruments in the cockpit show the pilots via highly ergonomic displays the attitude, height, climb rate, speed, and Mach number of the aircraft as well as the state of the engines and other factors such as the outside air temperature and the wind speed and direction. On-board weather radar informs the pilots of storms in the path with detailed information about the precipitation, turbulence, and the lateral and vertical extent of the storms. The navigation system takes inputs from GPS satellites, an inertial reference system (IRS), and VHF radio beacons; filters the information; and provides a precise indication of the position of the aircraft in three dimensions to within a few meters. These same instruments and navigation systems provide information to the autopilot, which can control the aircraft in height and position to follow a specific flight plan and land the plane at the destination airport if the latter has the necessary ground systems installed. The navigation computer contains a detailed database of all man-made and natural potential obstacles and provides warnings of approaching terrain or structures. The flight is conducted via a comprehensive air traffic control (ATC) system that tracks the aircraft and maintains communication links with the pilots throughout the flight to ensure safe separation with other aircraft. In addition, the aircraft will communicate automatically with others in the local area and build a three-dimensional map of all nearby flights to provide a traffic avoidance system that is independent of ground air traffic controllers. The system will not only warn the pilots of nearby traffic but also in extreme cases will inform them what evasive action to take. These and other systems have led to an unprecedented level of safety in commercial air travel which, if represented as fatalities per km traveled, is safer than any other type of transport on water or land [1]. This parameter does not

Aircraft Systems: Instruments, Communications, Navigation, and Control,
First Edition. Chris Binns.

Companion website: www.wiley.com/go/binns/aircraft_systems_instru_communi_Navi_control

necessarily provide the fairest comparison between different modes of travel since air travel will naturally do well with any safety assessment that uses distance traveled as the criterion. For example, when using fatalities per hour traveled, aviation drops to third on the list below rail and bus transport. It remains true, however, that air travel safety has made vast improvements by any measure in the last few decades and this is largely due to the incorporation of the systems listed above. All of these will be described in detail in subsequent chapters but before delving into the technical complexity it is worth exploring, in this chapter, a condensed history of the development of some of those instruments and systems.

1.2 The Advent of Instrument Flight

In the earliest days of aviation, the pilot's senses were the main aircraft instruments, with vision being used to estimate speed, height, and flight attitude while hearing and smell were used to monitor the state of health of the engine. The Wright flyer did have instruments installed including an anemometer, a stopwatch, and a revolution counter but these were used exclusively to analyze the performance of the flyer and the engine post landing. The flight itself was conducted entirely utilizing the senses of the pilot. Flying by senses alone dominated aviation throughout the First World War and led to the myth of instinctive balance when flying an aeroplane. This remained while flights were rarely conducted in bad weather and small rolled attitudes that developed inadvertently while flying through an individual cloud went unheeded. After the war, airmail and the first passenger services started to be developed but initially these flew under the weather sometimes at very low altitude resulting in many fatalities.

It was realized by the end of the First World War that flying in cloud with pilot vision completely removed from the available information could quickly lead to spatial disorientation and the aircraft spiraling out of the cloud with complete loss of control. The problem is that inner ear senses, which measure linear and angular acceleration, are evolved for life on the ground and provide misleading sensations in aircraft. Examples include the Somatogyral illusion in which an established banked turn is undetected as there is no angular acceleration but rolling out of the turn produces the illusion of a bank in the opposite direction, and the Somatogravic illusion where accelerations and decelerations are interpreted as pitches up and down, respectively. Gyroscopic turn coordinators were available by 1918 but without instrument training pilots still tended to favor their senses over the indications of the instrument.

Early pioneers in the development of instrument flight were two US army pilots, William Ocker and Carl Crane. By 1918, the first gyroscope-based

attitude indicators (AIs) (see Section 1.4 and Section 3.1.9), invented by Elmer Sperry, were available and Ocker was one of the first to attempt an extended flight in cloud using the instrument. The flight still ended up with the aircraft in a spiral dive but Ocker realized that the main reason was his failure to put complete faith in the instrument and to pay too much attention to his erroneous balance senses. Ocker was one of the first to correctly identify the misinformation coming from balance organs and became somewhat of an evangelist for using instruments in flight. Crane was nearly killed in 1925 when he dropped into a spiral dive out of cloud while flying a congressman's son to Washington and was acutely aware of the problems of maintaining control while blind. Ocker and Crane teamed up in 1929 and conducted a comprehensive study of flying in clouds, which led, in 1932, to the publication of their book, *Blind Flight in Theory and Practice*, which is the first systematic exploration of instrument flight. By the late 1920s, a full range of pressure and gyro instruments were available as well as some radio navigation devices (see below) and in 1929 Jimmy Doolittle demonstrated a "blind" takeoff, aerodrome circuit, and landing in an aircraft whose dome was covered [2]. The cockpit in Doolittle's NY-2 Husky biplane is shown in Figure 1.1a and contains the six main glass instruments that are to be found in a current general aviation (GA) light aircraft. That is, an altimeter (i), an AI (ii), an airspeed indicator (iii), a turn indicator or turn coordinator (iv), a direction indicator (DI) (v), and a vertical speed indicator (VSI) (vi). By the 1950s, the layout of these six instruments was standardized into what was deemed to be the most ergonomic arrangement (the so-called "6-pack") and they were mounted as shown in Figure 1.1b, which shows a Piper PA28 cockpit.

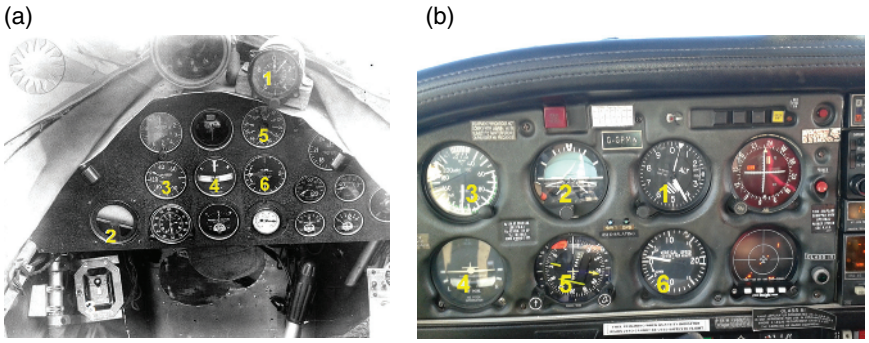


Figure 1.1 (a) Flight instruments in the cockpit of the NY-2 Husky biplane used in the first "blind" takeoff and landing flight by Doolittle in 1929. *Source:* Reproduced from Ref. [2] with permission of ETHW. (b) The same six instruments in the standard layout in a 1960's light aircraft (Piper PA28).

In older large aircraft with traditional instruments, the standard 6-pack is also evident directly in front of the pilot though it is embedded in an extended array of engine and navigation instruments. The main change to this layout came in the transition to "glass cockpits" in the late 1960s where several instruments are displayed on a single electronic screen. The term is slightly misleading as there is probably less glass in a glass cockpit than a traditional one with a large array of glass-fronted instruments but basically it means information is displayed on electronic screens rather than individual instruments. The change to glass cockpits marked the transition from direct-sensing to remote-sensing instrumentation. In the case of older direct-sensing pressure instruments, the pressure being measured is brought via tubes directly into the back of the instrument, which then converts it into a reading on the instrument face as described in Chapter 2. This leads to a large amount of tubing mixed in with all the wiring behind the instrument panel. In remote sensing, a transducer measures the quantity required remotely and converts it into an analog or digital electrical signal, which is conveyed by wires, either to an individual instrument, or to a computer and display generator. In the most modern systems, a digital data bus is used to convey information from all the sensors, which significantly reduces the complexity of the wiring. Figure 1.2 compares a cockpit with traditional direct-sensing instruments in a twin piston engine aircraft (Figure 1.2a) and a glass cockpit in which the remote-sensing instruments communicate via a computer to the electronic displays, again in a twin piston engine aircraft (Figure 1.2b). Note, however, that even in the glass cockpit there are some direct-sensing instruments provided as backup in case of a total power failure. The glass screen immediately in front of the left pilot seat is referred to as the Primary Flight Display (PFD).

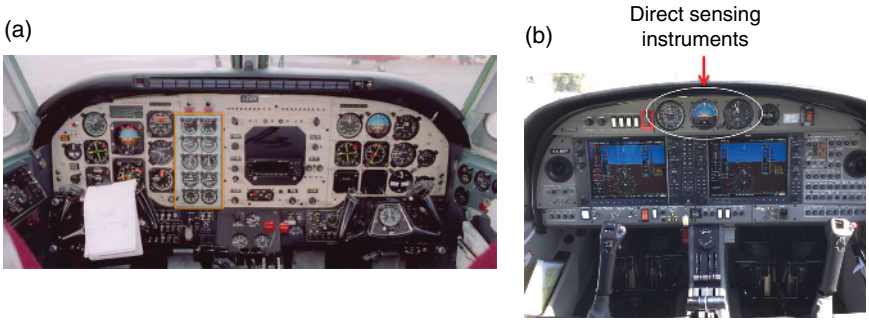


Figure 1.2 (a) Instrument display in the cockpit of a twin piston engine aircraft using entirely analog direct-sensing instruments. (b) Instrument display in the cockpit of a twin piston engine aircraft utilizing remote-sensing instruments and digital displays connected by a digital data bus (glass cockpit). Note, however, that some direct-sensing analog instruments are also provided as backup.

An important aspect of instrument flight is training and although the technology for flight without external references was in place by 1930, pilot training in instrument flying was not standardized internationally until after the International Civil Aviation Organisation (ICAO) was set up in 1947. The modern day "instrument rating" is a separate rating applied to the pilots' license allowing the holder to fly on instruments only and in addition to navigate and land using radio navigation aids (see Chapter 7). This chapter will now describe the historical development of individual instruments and also the evolution of the communication and navigation systems essential to modern aviation.

1.3 Development of Flight Instruments Based on Air Pressure

1.3.1 The Altimeter

Measuring how high an aircraft is off the ground became a necessity as soon as flights over high ground started to become commonplace. Since the beginning of aviation, a number of methods have been tested including sonar, variation in gravity, capacitance, integrating accelerometers (IRSs), cosmic ray detection, and hypsometry (measuring the change in the boiling point of water). The method that became established early on and continues to this day is to measure the air pressure and convert this to an altitude. The French physicist Blaise Pascal first confirmed the decrease of air pressure with increasing altitude in 1648 using a mercury barometer invented by Torricelli four years earlier. He measured the pressure at the bottom and top of a church bell tower in Paris and was able to observe a measurable decrease produced by climbing 50 m to the top. Soon after the first balloon flight by the Montgolfier brothers in 1783, portable mercury barometers were being used to estimate altitude in free balloons, but practical altimeters were not available till Bourden produced an improved design of aneroid barometer in about 1845. From then on altimeters based on the Bourden design were commonly used in free balloons and airships.

The Wright flyer did not have an altimeter and strangely there is no confirmed record of altimeters in use in aircraft before 1913. The first operational aircraft altimeters, available from about 1912, had a single pointer that completed one revolution in 0–10 000 ft (Figure 1.3a). By about 1925, the altitude range had been extended to 0–30 000 ft, still employing a single pointer but in this case, it completed one and a half revolutions with an inner scale used to read altitudes above 20 000 ft. At about the same time, adjustable scales were introduced so that the altimeter could be adjusted for changes in air pressure that occur day to day and also for differences in elevation between departure and arrival aerodromes. The development of radio links (see Section 1.5) enabled altimeters to be set for the measured pressure at a landing field and by the late 1920s,

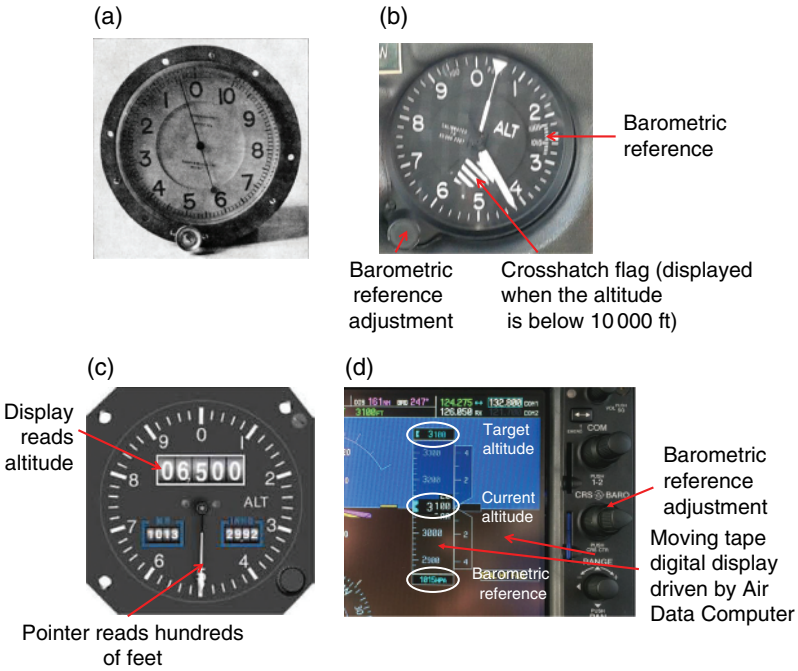


Figure 1.3 Historical evolution of altimeters (a) Circa 1912, single pointer and a range of 10 000 ft. *Source:* Reproduced from Ref. [3]. (b) Three-pointer display seen after 1935 and still in common use today. The large, small, and thin line pointers indicate hundreds, thousands, and tens of thousands of feet, respectively. (c) More modern presentation in which the barrel is a digital display of altitude and the pointer reads hundreds of feet. (d) Moving tape display driven by the Air Data Computer in a glass cockpit PFD.

temperature compensation was built in, which made altimeters accurate enough to be useful for landing. In order to produce a high resolution but still maintain a large altitude range, three-pointer altimeters were introduced in 1935 (Figure 1.3b). In these, the largest pointer makes one revolution in 1000 ft pointing to a number that represents hundreds of feet while a smaller pointer moves one revolution in 10 000 ft pointing to a number that represents thousands of feet. The third hand, which is usually a thin line, indicates tens of thousands of feet. This type of altimeter is still found in most light aircraft flying today. A major improvement was the introduction of the servo-altimeter in which the aneroid capsule used to measure pressure does not directly drive the indication mechanism, but its distortion is measured electrically and this electrical signal drives the display (see Section 2.8.3). More recently, the problem of the easy misreading of the three-pointer dial has resulted in the display shown in Figure 1.3c in which the altitude is displayed directly on a digital barrel

and the pointer indicates hundreds of feet with 20 ft divisions. In modern commercial airliners and GA aircraft with a glass cockpit, the air pressure is measured by a transducer and the electrical signal along with other data such as the outside air temperature is passed into an air data computer (ADC). The output of the ADC is then used to drive a digital moving tape-type display shown in Figure 1.3d). Although measurement of air pressure remains the primary method of determining altitude, Radio Altimeters (see Section 5.6) are the primary input into the autopilot during landing. In addition, now that accurate models of the surface of the Earth are available (see Section 6.4), satellite and inertial reference systems can also provide accurate altitude information.

1.3.2 The Vertical Speed Indicator (Variometer)

As in the case of the altimeter, the history of the VSI begins before the advent of heavier than air flight. Balloonists used instruments called *statoscopes*, which were sensitive aneroid barometers to detect climb or descent though they did not indicate the rate. In the twentieth century, gliders drove the development of the VSI or *Variometer*, as it is referred to in gliding, because of its importance for finding optimum lift in updrafts. The human body is insensitive to small gradual changes in barometric pressure that are typical of normal climb rates so to optimize endurance and range a variometer is critical. The earliest variometers (Figure 1.4a) were based on statoscopes but later designs measured directly the rate of change of air pressure to provide a quantitative reading. The gliding pioneers Alexander Lippisch and Robert Kronfeld are credited with the invention of the first truly quantitative variometers in 1929, which greatly improved glider performance. Their design was based on having a sealed

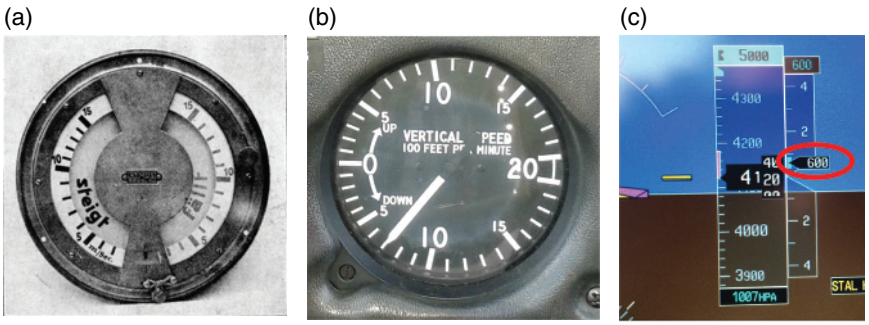


Figure 1.4 Development of the Vertical Speed Indicator (VSI). (a) Atmos variometer from circa. 1922. *Source:* Reproduced from Ref. [3]. (b) Direct reading VSI found in a General Aviation aircraft showing a 700 ft min^{-1} rate of descent. (c) Digital indication and pointer (highlighted by the red oval) next to the altitude moving tape in a glass cockpit PFD showing a 600 ft min^{-1} climb.

container connected to a diaphragm, the other side of which was at ambient pressure. As the air pressure changed, air would flow into or out of the container and the rate of height change was determined by the flow rate.

The current design of direct reading VSIs used in GA (see Section 2.9) is based on an aneroid capsule whose internal part and surround are both connected to ambient air but a controlled leak between the inside and outside of the capsule maintains a pressure differential if the air pressure is changing. The first report of this design was by Wing Commander Roderic Hill in his book *The Baghdad Air Mail* [4] first published in 1929. He described how hard it was to climb in the Vickers Vernon biplanes used for the mail runs in the Middle East when they were taking off fully laden from hot aerodromes well above sea level. The pilots resorted to using updrafts (termed *dunts*) to climb the first few hundred feet and initially would try and find them by observing birds. One of them built a *dunt indicator* by drilling a small hole in a two-gallon petrol can and connecting the can to one side of a pressure gauge the other side of which was exposed to ambient air. If the pressure changed due to a climb or descent, because of the time lag required for the pressure inside the can to equalize, the gauge would indicate a pressure differential. A current design of direct reading VSI is shown in Figure 1.4b.

More recently, the different requirements of powered flight and gliders has led to a divergence in the design of VSIs and variometers. The most important quantity to a glider pilot is the change in the total energy (potential plus kinetic) and modern glider variometers have total energy compensation. Thus, they distinguish whether climb is coming at the expense of kinetic energy, which would happen, for example, by simply applying back pressure to the stick. In powered flight, on the other hand, the pilot needs to know the absolute rate of climb or descent irrespective of speed. In a glass cockpit, the vertical speed indication comes directly from the ADC (see Section 2.14), which analyzes digitally the pressure measured by the static source and determines if there is an upward or downward trend. The vertical speed indication is normally given by a pointer and a digital display next to the altitude moving tape as shown in Figure 1.4c.

1.3.3 The Airspeed Indicator

The *true airspeed* (TAS) of an aircraft is defined by its speed relative to the still air around it that is close enough to be in its immediate environment but sufficiently distant to remain undisturbed by its passage. Thus, it is the speed of the aircraft relative to something invisible and all devices that determine airspeed measure the force generated by the pressure of the passing air on some surface, that is, the *dynamic pressure*. The direct measurement of dynamic pressure leads to a quantity known as the *indicated airspeed* (IAS), which will vary relative to the TAS depending on the air density around the aircraft. For example, if the aircraft climbs at a constant IAS, that is, maintaining a constant dynamic

pressure at the measurement device, its TAS will steadily increase as the air gets thinner. The relationship between TAS and IAS is described in detail in Section 2.10, but given that IAS almost never equals TAS it may seem surprising that it is always IAS that is primarily presented by the cockpit instruments to the pilot. In addition, the operating handbooks of aircraft give speeds for safe operation of procedures such as lowering flaps or landing gear in terms of the IAS. The reason is that the IAS is essentially a measure of dynamic pressure and all the important flight parameters, such as lift, stall conditions, critical speeds for lowering landing gear and flaps, etc. depend on the dynamic pressure, irrespective of the TAS. Thus, the IAS can be used by the pilot to maintain safe flying conditions without having to worry about converting to TAS. The latter is only important to determine ground speed (after correcting for wind), which is required for navigation. In transonic and supersonic aircraft, the Mach number is another important indication required for safe flight, which in older aircraft is measured from the static and dynamic pressure by a separate instrument known as a Mach Meter (see Section 2.11). In more modern aircraft, the Mach number is calculated by the ADC and displayed on the PFD.

The earliest airspeed indicators were anemometers, similar to those used to measure wind on the ground, attached to a revolution counter as illustrated in Figure 1.5a or a metal plate perpendicular to the airflow on a spring (Figure 1.5b).

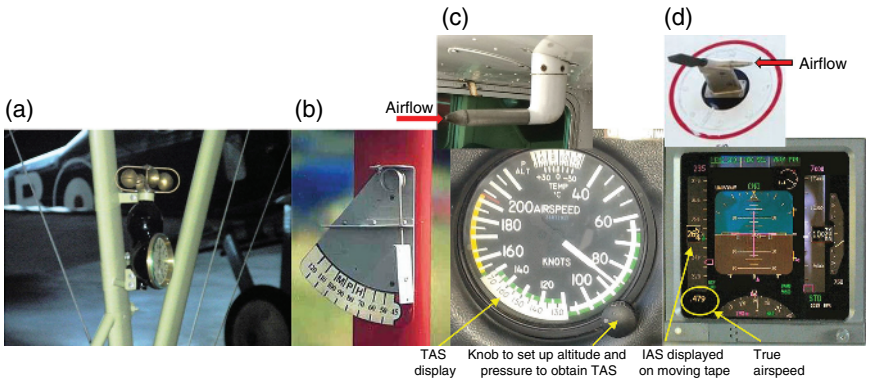


Figure 1.5 Evolution of airspeed indicators. (a) Anemometer-type airspeed indicator on the wing strut of a WW1 Albatross D.V. biplane. *Source:* Reproduced with permission of 20thcenturybattles.com. (b) Plate on a spring used to indicate airspeed on the wing strut of a De Havilland 60 Moth (1930s). *Source:* Reproduced with permission of James Knightly. (c) Pitot tube and direct reading Airspeed Indicator (ASI) on a General Aviation aircraft. (d) Pitot tube mounted on a rotating table on the fuselage. *Source:* Reprinted under Creative Commons license 3.0 [5] and PFD on a current commercial aircraft (Boeing 737NG). The moving tape indicates the IAS and the highlighted digital display shows the TAS or Mach number. *Source:* Reproduced with permission from Chris Brady from http://www.b737.org.uk.

By the end of the First World War, the majority of aircraft used a pitot tube, that is, a closed tube in which air is brought to rest (Figure 1.5c). By measuring the difference between the *stagnation pressure* in the tube and the static air pressure (i.e. the *dynamic pressure*), it is possible to determine the airspeed. The pitot tube is a very simple measuring device, originally invented in 1732 by Henri de Pitot to measure water flow and improved on by Darcy in 1856. It is still in widespread use and has remained fundamentally unchanged though improvements have been made by adding internal heaters to prevent icing and installing them on rotating platforms with vanes so that they are always presented in the same direction to the airflow (Figure 1.5d).

In direct-reading airspeed indicators (Figure 1.5c), the pressure difference generated by the pitot tube is sensed by an aneroid capsule whose distortion drives the indication as described in Section 2.10. In aircraft with glass cockpits, the stagnation pressure in the pitot tube and the static pressure are measured by transducers and the signals are processed by the ADC (see Section 2.14), which outputs the IAS to a moving tape on the PFD as illustrated in Figure 1.5d. The ADC has all the other necessary parameters (outside air temperature, altitude, etc.) to calculate the TAS or Mach number and this is also displayed on the PFD as highlighted in Figure 1.5d.

1.4 Development of Flight Instruments Based on Gyroscopes

The properties of gyroscopes and their use in aircraft instruments are described in detail in Section 3.1, but a brief introduction is given here. One of the most important characteristics for their use in instrumentation is the tendency of a spinning rotor to maintain the same spin axis (its *rigidity*). The observation of the stability of a spinning top must precede recorded history and the earliest examples discovered are clay tops from the Middle East dating back to 3500 BCE. In order to make use of this stability for orientation and build a practical gyroscope, however, the spinning top (or *rotor*) needs to be mounted in gimbals as shown in Figure 1.6. In the case of the rotor mounted in a simple cage (Figure 1.6a), the cage can rotate freely about the spin axis but torque applied about any other axis will be resisted and generate a torque about a perpendicular axis (see Section 3.1). If the cage is mounted in a gimbal that pivots about a perpendicular axis (Figure 1.6b), the support structure can now rotate freely about two axes without disturbing the rotor. Adding a second gimbal (Figure 1.6c) allows the support structure to rotate about all three axes so if the support is clamped to the frame of an aircraft the rotor will maintain its spin axes irrespective of changes in the pitch roll or yaw. If the gyroscope rotor is initially aligned with the horizon, for example, its rigidity will always indicate

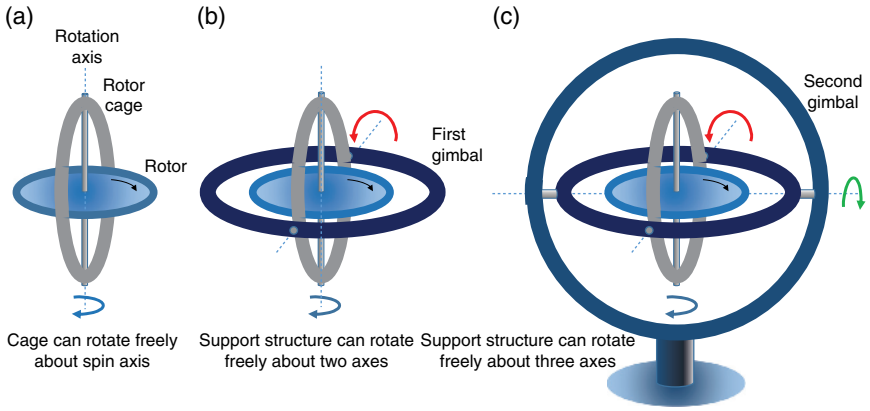


Figure 1.6 (a) In a simple caged rotor the cage can rotate freely about the spin axis without applying torque to the rotor. (b) After adding a single gimbal the support structure can rotate about the two axes shown without applying torque to the rotor. (c) After adding a second gimbal the support structure can rotate about all three axes without applying torque to the rotor. The outer gimbal can be clamped to the frame of an aircraft and the spinning rotor will maintain its axis of spin as the aircraft maneuvers.

the plane of the horizon as the aircraft maneuvers. How this characteristic is used to build a practical instrument to show the orientation of the horizon, that is, the *artificial horizon* or AI is described in Section 3.1.9.

The first reported use of a gyroscope in navigation was in 1743 by the English sea captain John Serson who invented his *whirling speculum* to locate the horizon on misty mornings – an early form of artificial horizon. The name gyroscope (from the Greek *gyros* for circle and *skopeein* to observe) was first introduced by Léon Foucault who used one in 1852 to demonstrate the rotation of the Earth. The German inventor Hermann Anschütz-Kaempfe patented the first practical gyrocompass in 1904 and an improved design was presented by the American entrepreneur Elmer Sperry later that year. Initially, gyroscopes were used in ships but the Sperry Gyroscope company, founded in 1910, pioneered the use of gyroscopes in aircraft instrumentation. Lawrence Sperry, Elmer's son, used gyroscopes to build the first functional autopilot prior to World War 1, as described in Section 1.9. The first artificial horizon was installed in an aircraft in 1916 and by the time Doolittle made the first blind instrument flight in 1929 (see Section 1.2) the full set of gyro instruments found in a modern 6-pack, that is the DI, the turn indicator, and the AI were all available. These remain in widespread use today, largely unchanged, in most GA aircraft and some older commercial airliners.

The introduction of glass cockpits in the 1960s changed the display of information given by gyroscopes from individual direct reading instruments to the

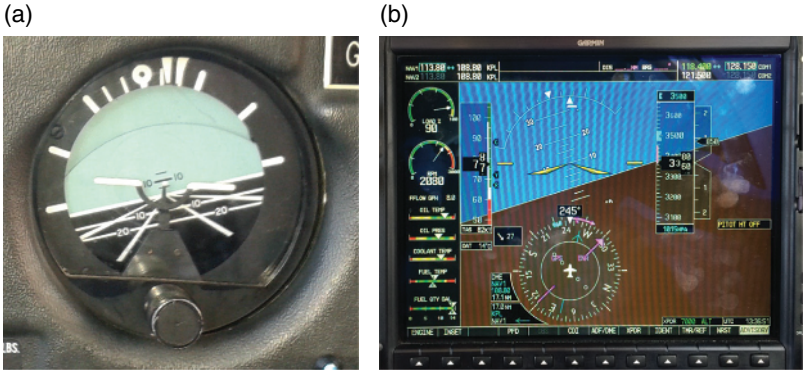


Figure 1.7 (a) Roll and pitch indications in direct reading instruments in a general aviation aircraft performing a climbing turn (10° pitch up, 15° bank). (b) Roll and pitch indications on the PFD in an aircraft with a glass cockpit performing a climbing turn (10° pitch up, 15° bank).

multi-instrument PFD but the spinning mechanical gyroscopes were still used to determine the attitude information. This all changed with the introduction of Micro-Electrical–Mechanical Systems (MEMS) gyroscopes in the early twenty-first century. These are based on the rigidity of the plane of vibration of a tuning fork (see Section 3.2.5) and their main advantage is that they can be micromachined into silicon wafers so that a three-axis gyroscope can now be mass-produced cheaply on a piece of silicon a few millimeters across. On the same chip can be mounted a number of other devices as described in Section 3.4 so that an entire attitude and heading reference system complete with gyroscopes, magnetometers, and accelerometers can be mounted on a single chip. Apart from savings in cost and weight, MEMS devices are more reliable than mechanical rotating gyroscopes. Figure 1.7 shows the AI display found in the traditional "6 pack" of direct-reading instruments in GA aircraft performing a climbing turn and the same information displayed on the PFD in a glass cockpit.

Gyroscopes are also used in IRSs, which were initially developed in rockets starting from the 1930s and then were widely incorporated into aircraft navigation systems post World War 2. Initially, these were also based on mechanical spinning gyroscopes but from the 1970s the mechanical gyros were replaced with laser-based systems (see Section 3.2).

1.5 Development of Aircraft Voice Communications

In the earlier days of aviation, communication between pilots and people on the ground was either by hand signals, aircraft maneuvers (rocking wings, etc.), or

by large symbols on the ground. Interestingly, all three forms of communication have survived in aviation into the twenty-first century, having been standardized internationally and are still in the training syllabus for commercial pilots. Hand signals survive in the form of a set of hand signals for communication between pilots and ground marshallers. Large symbols prepared on the ground following a plane crash have been devised to communicate with search and rescue (SAR) aircraft. Many GA airfields still have a "signals square," which is a set of large symbols visible from the air that communicate information about the airfield and the runway in use. These are for use either by vintage aircraft with no radio or by aircraft that have suffered a radio failure. In the case of loss of radio communication at night, signals using aircraft lights can be used. Finally, a series of aircraft maneuvers for communication between commercial airliners and intercepting military aircraft has been standardized.

The world's first radio transmission from an aircraft was a Morse code message sent in August 1910 by the Canadian aviator, James McCurdy, as he flew over a receiver in Brooklyn, New Jersey, trailing a long aerial wire from his aircraft. Within a few weeks a Morse message from air to ground was transmitted in excess of one mile in England. Leading into the First World War, the Marconi company in collaboration with the Royal Flying Corps began a research program on air to ground and ground to air communications, which was driven by the need to communicate observations from the air to artillery on the ground. Prior to radio communication this had been achieved by dropping hand-written notes from the aircraft. Initial attempts at wireless transmission used Morse code and significant problems were soon encountered such as the weight of the early transmitters – 35 kg or more plus 250 ft or so of aerial wire that had to be unwound from a spool. This created significant drag and sometimes the wire would get caught around the control surfaces creating a serious hazard. A plane was a death trap if it was attacked while set up for transmission. In addition, it was very hard for pilots to tap out a message on a Morse key while flying the plane. Things were not much better for ground to air transmissions as it was found that hearing a Morse message above the roar of the engine and the wind and sometimes gunfire was challenging to say the least.

The goal soon became to send voice transmissions over a significant range and the first step in this quest was achieved in April 1915 when Captain J.M. Furnival allegedly heard the following message in his headphones from Major Prince while flying over Brooklands Park in England:

> *Hello Furnie. If you can hear me now it will be the first time speech has ever been communicated to an aeroplane in flight.*

Within a few months, voice communication in both directions was achieved and Marconi started manufacturing the world's first production aviation radio weighing just 9 kg. Across the Atlantic, the technology was developing at the

Figure 1.8 AT&T employees and military personnel watch an early aircraft-radio test. *Source:* Courtesy of the AT&T Archives and History Center.

same rapid pace and by July 1917, AT&T engineers demonstrated two-way voice communication at Langley Field in Virginia (see Figure 1.8). By August of that year they had also produced two-way communication between aircraft in flight.

The radio communications developed during the First World War were specific to the requirements of the aircraft used to observe the position of landing artillery shells and the messages were transmitted over quite short ranges. The radios of the time worked in the Low Frequency (LF) band (30–300 kHz) in which the radio waves are refracted by the ground and tend to follow the curvature of the Earth (see Section 4.2.2.1). The maximum range in this band is given by the transmitter power and for practical transmitters that could be carried on aircraft the range was limited to about 20 miles. After the war, the requirements changed and the ability to communicate with aircraft over a long range became important as the first air mail routes were established.

Initially, messages were bounced from airfield to airfield leading to anecdotes of aircraft arriving at a destination airfield before the radio message that it was on its way had been received. In the longer term the range of radios had to be

Figure 1.9 Imperial Airways Short S23 *Caledonia* flying boat at Felixstowe in 1936. In July 1937, *Caledonia* flew from the Foynes seaplane port in Ireland to Botwood, Newfoundland maintaining radio contact with Foynes for the entire 1900-mile journey.

increased and this was achieved by developing transmitters working in the Medium Frequency (MF: 300 kHz to 3 MHz) and High Frequency (HF: 3–30 MHz) bands. At these frequencies, it is possible to utilize the ionosphere to reflect radio waves around the curvature of the Earth (so-called *skywaves* – see Section 4.2.3) producing very long ranges. By 1932, Marconi had developed the AD37/38 radio that used both the MF and HF bands and could achieve two-way voice transmission with the ground at a range of 1000 miles and communication between aircraft at a range of 200 miles. In addition, the radio could be used as an Automatic Direction Finder (ADF: see Section 1.7.1 and Section 7.1). During the 1930s radios transmitting in the MF/HF band continued to improve and became the norm for aircraft communications in the rapidly developing commercial aviation sector throughout the 1930s. In July 1937, an Imperial Airways Short S23 flying boat (*Caledonia*, Figure 1.9) took off from the Foynes seaplane port in Ireland and flew to Botwood in Newfoundland maintaining radio communication with Foynes for the entire 1900-mile flight.

Thus, at the start of World War 2, MF/HF long distance radios were highly developed for commercial air travel and were also used in military aircraft. It was decided that improvements were required in the performance, frequency range, and ease of use and so in 1939, new specifications were drawn up and presented to Marconi who developed the T1154/R1155 transmitter/receiver combination. The transmitter had a continuous wave output power of up to 70 W and the entire combination weighed about 36 kg. In addition to communication, the installation could be used in an ADF mode providing a radio navigation capability. Bomber command had the radio sets installed in their aircraft by June 1940 (see Figure 1.10), and by the end of the war over 80 000 sets had been manufactured. After the war, many sets found their way into civilian planes

Figure 1.10 T1154/R1155 installation at the radio operators station in an Avro Lancaster, circa 1943.

and they continued to be used in military aircraft (Vickers Varsity and Handley Page Hastings) till these were withdrawn in the 1970s.

The T1154/R1155 installation was fine for large planes but was not suitable for fighters because, apart from its size, the MF/HF frequency bands used presented a more fundamental problem. Generally, the lower the frequency (i.e. longer the wavelength) of radio waves used, the longer the antenna needs to be (see Section 4.4). The Lancaster had two antennas – a fixed one for HF transmission running from above the cockpit to the tail and a trailing 290-ft cable that could be deployed as an MF antenna. A trailing antenna was not an option for a fighter and the relatively short fixed antennas that could be used were inefficient in the HF band severely restricting the range. In addition, the HF band is prone to a lot of interference resulting in a high level of static noise in the crew's headphones. It was decided at the beginning of the war to equip Spitfires and Hurricanes with VHF (Very High Frequency: 30–300 MHz) radio installations [6] so that an efficient short antenna could be used and transmissions would suffer less static noise. The result was the Marconi TR1133 and subsequently TR1143 transmitter/receiver in a single box weighing 21 kg and with a transmitter power of 10 W. Radio waves in the VHF range are not reflected by the ionosphere, nor do they follow the curvature of the Earth due to refraction like LF waves so the range is limited by the distance to the horizon and thus the range depends on the height the aircraft is flying. As shown in Figure 1.11, the distance, d, from an aircraft to the horizon is one side of a right-angle triangle

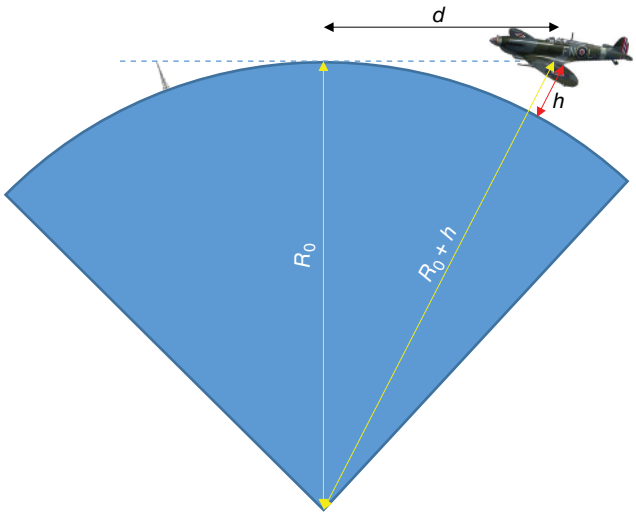


Figure 1.11 Line of sight distance, *d*, of an elevated object to the horizon forms one side of a right-angle triangle whose other sides are R_0 and $R_0 + h$, where R_0 is the radius of the Earth and *h* is the height of the object. To get line of sight to a second elevated object the two line of sight distances are added.

whose other sides are of length R_0 and $R_0 + h$, where R_0 is the radius of the earth and h is the height the aircraft is flying. Thus, we can write:

$$d^2 + R_0^2 = (R_0 + h)^2$$

and since h^2 is insignificant compared to R_0^2 this becomes:

$$d = \sqrt{2R_0 h}$$

Normally we would want to enter the value of h in feet and calculate d in nautical miles. So, putting in the value for $R_0 = 3440$ nautical miles and converting d from feet to nautical miles, that is dividing it by 6076.12 gives:

$$d(\text{nm}) = 1.064\sqrt{h(\text{ft})}$$

This is always increased by about 15% to take into account the fact that VHF waves can be picked up slightly beyond the horizon giving:

$$d(\text{nm}) = 1.23\sqrt{h(\text{ft})}$$

(some sources specify the pre-factor as 1.25). This also assumes the ground station antenna is at ground level but if the antenna has a significant elevation or the aircraft is talking to another aircraft then it is simply a matter of adding the

second station's line of sight distance to the horizon, so for two elevated objects, the range can be written:

$$d(\text{nm}) = 1.23\left(\sqrt{h_1(\text{ft})} + \sqrt{h_2(\text{ft})}\right) \tag{1.1}$$

where h_1 and h_2 are the heights of the two communicating entities. As an example, even at 2000 ft, the range is about 55 nautical miles to a ground station with no elevation and double that for two aircraft at the same height, which was adequate for normal fighter operations.

Thus, emerging from World War 2, aircraft communications had evolved into dual modalities with VHF being used for high fidelity short-range communications and HF band skywave transmissions for long-range messages. Remarkably, these two types of transmission are still in place 70 years later and the main evolution in the equipment has been due to advances in electronics. From the late 1950s, transistors started to replace valves and from the mid-1960s integrated circuits replaced individual transistors and other circuit components. Integrated circuits have continued to evolve with increasing levels of integration from about 10 transistors on a chip in the mid-1960s to over seven billion in 2015. Note, however, that it is the control aspects that can be handled by integrated circuits while the main transmitter power still has to be derived from individual power transistors. Nevertheless, a typical transmitter box on a GA aircraft with a similar power to the TR1143 unit in a spitfire weighs less than 2 kg and would fit in a coat pocket. In addition, it is multichannel with both communication and VHF radio navigation capabilities (see Section 1.7.3 and Section 7.2).

VHF radio transmissions (at similar frequencies to those used in World War 2) remain the main communication mode for air to ground and air to air for all terminal operations and also for traffic over land, with aircraft switching to different stations as they travel to stay in range. For trans-Oceanic flights, VHF does not provide coverage for most of the flight and from World War 2 till the mid-1970s this gap continued to be filled by HF skywave communications but the disadvantages of this started to become increasingly apparent. The optimum frequency to use to achieve long ranges depends on the state of the ionosphere, which varies considerably in the diurnal cycle, so using HF equipment properly requires significant training. The reliability of transmission is also at the mercy of natural phenomena such as sunspot cycles and solar flares (so-called *space weather*) leading to the setting up of space weather monitoring services to try and predict periods when HF communication is likely to be interrupted. From an operational point of view, there is a high background noise level on HF radio frequencies and listening for transmissions for long periods of time was tiring for aircrews. This led to the setting up of the SELCAL system where each aircraft has a four-letter code, which, for a trans-oceanic flight, is included in its flight plan. If a ground station operator wants to communicate with a

specific aircraft, they will transmit the corresponding code, which all aircraft on frequency will receive but the one that has been selected will activate a chime alerting the crew that a message is coming in. Thus, the crew can turn down the volume of their headsets until an alert is received.

1.6 Development of Aircraft Digital Communications

1.6.1 Communication Via Satellite (SATCOM)

The SELCAL system used with HF communication is an early example of digital communication as opposed to voice, but as aviation moved into the digital age the main problem with HF communication was its inability to transmit data at a high rate. For transmission of digital data, the carrier wave needs to be at a higher frequency than the modulation frequency and a high data rate demands a very high carrier wave frequency, which restricts the range to line of sight. From the early 1970s, this range limitation of VHF transmissions was removed by using satellites. A satellite orbiting at a height of 22 236 miles (35 786 km) above the Earth's equator is *geostationary*, that is, it has the same angular velocity as the spinning Earth below and so it stays fixed at one point in the sky as seen from the surface. It has line of sight with everything below and a constellation of a few geostationary satellites can cover the globe. A schematic of the system used to communicate between aircraft and ground via satellites is shown in Figure 1.12. An operator who can be an air traffic controller or from airline operations staff will send a message to the network run by the satellite operator (for example, Inmarsat), which will provide not only voice communication but digital services such as the Aircraft Communications Addressing and Reporting System (ACARS: see Section 1.6.3) or the internet. The message (along with all the other services) is sent to a ground station satellite dish, which is pointing permanently at the geostationary satellite. Data and voice are transmitted to the satellite via a Super High Frequency (SHF: 3–30 GHz) band signal and then relayed by the satellite to the aircraft in flight via an Ultra-High Frequency (UHF: 300 MHz to 3 GHz) link. This mode of communication is now the norm and its capabilities are constantly expanding to include, for example, monitoring of aircraft in flight via ACARS. The need for ever higher data transmission rates is pushing the radio frequency higher and in the latest system under development this will reach 40 GHz, corresponding to 1 cm waves normally used in radar.

Hence, is HF skywave communication now redundant? Far from it. Satellites have their own problems, for example, their inaccessibility for repairs if they fail. Most airlines maintain the capability for HF transmission as a failsafe backup to their normal satellite services for long-range communications and the system infrastructure has been continued. This has been proven to be a wise policy,

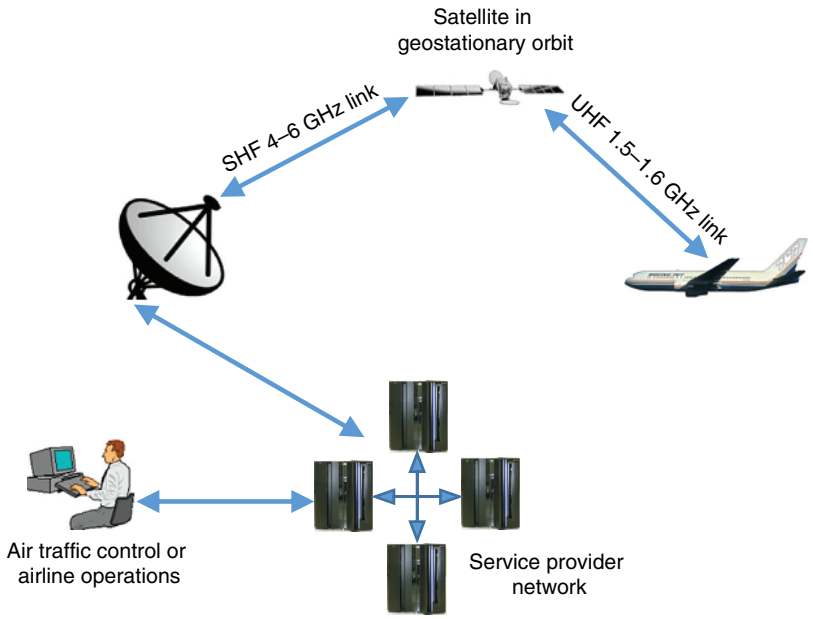


Figure 1.12 Typical system for satellite communication (SATCOM). The operator (ATC or airline staff) sends a message to the service provider network that also transmits digital data (e.g. ACARS or internet). The message is sent to a ground station whose dish is aligned with the satellite in geostationary orbit. The ground station transmits the voice message and data via an SHF link and the satellite relays everything to the aircraft in flight via a UHF link.

especially in the case of emergencies and natural disasters. There are numerous instances including the 2004 Indian Ocean tsunami, in which damaged or overloaded satellite and telephone systems have meant that humanitarian and emergency agencies have had to make extensive use of HF radio to coordinate their missions. In addition, the polar regions still have unreliable satellite coverage making it necessary to use HF radio communication. The HF communications equipment in aircraft and the ground infrastructure required to support it will stay for the foreseeable future.

1.6.2 Secondary Surveillance Radar (SSR) and Traffic Alert and Collision Avoidance System (TCAS)

Primary radar was first developed in the mid-1930s and by 1936 a chain of radar stations along the South and east coast of England called the Chain Home system (Figure 1.13) was installed in time for the outbreak of World War 2. Primary radar works by emitting a pulse of radio energy along a specific direction and measuring the time taken for the reflected pulse from an aircraft to

Figure 1.13 Three of the transmitter masts of the Chain Home system photographed in 1945. The antenna itself is the wire array just visible to the right of the masts.

return giving the distance along that direction and hence the aircraft position (see Section 5.1). Problems that were recognized immediately from the birth of the technology included the weakness of the return pulse and the lack of other information about the aircraft including whether it was friend or foe. The original radar patent granted in April 1935 [7] contained an idea for overcoming these problems by having a transmitter installed on the aircraft that would detect the primary pulse from the radar and be triggered into transmitting its own signal. This would greatly increase the strength of the return signal, which could also contain information to identify friend from foe (IFF). The device on the aircraft was called a transponder (combination of transmitter and responder) and the system was known as Secondary Surveillance Radar (SSR: see Section 5.4). SSR was also developed independently in Germany and the United States, and by 1939 all three countries had their own IFF systems for aircraft and ships. Originally, transponders responded with a radio signal that had the same frequency as the primary pulse but as radar frequencies changed, it became necessary to transmit the signal in a separate band and this was set at 1000 MHz, which is very similar to the frequency used by transponders today.

After the war, the system was adapted for civilian use and in the first manifestation of civil SSR (Mode A), an aircraft transponder would transmit a four-digit octal *squawk* code set by the pilot to the radar station to help identify it. The term squawk originates from the codename "parrot," which was used when discussing the secret IFF system and is now a permanent part of aviation nomenclature. Another problem with primary radar is that it can only measure the slant distance and apart from nearby traffic is unable to determine the altitude. From the mid-1960s a second mode (Mode C) was introduced into SSR that would transmit a second four-digit octal code in addition to the squawk code that would report the aircraft altitude. It was soon realized that it would be possible for transponders to transmit much larger quantities of digital data and still maintain compatibility with both Mode A and Mode C and from 1990 a new mode called Mode S (for "Select": See Section 5.4.5) started to be installed.

Mode S eliminates some of the problems encountered with Modes A and C caused by overlapping signals in regions of heavy traffic and can transmit a code that is unique to a specific aircraft (assigned to it at manufacture). In addition, it not only communicates with the ground station but also with other aircraft and can interrogate the transponders of nearby traffic. Research into autonomous collision avoidance systems has been conducted since the 1950s and was given particular impetus in 1956 when a Lockheed Super Constellation and a Douglas DC7 collided over the Grand Canyon killing 128 people. Incorporating a directional SSR antenna so that the position of other traffic could be determined meant that Mode S SSR with its ability to automatically communicate with other aircraft provided a good platform to realize this, which led to the birth of TCAS. There have been various implementations of TCAS since its inception and the current system is TCAS II, version 7.1 but upgrades are planned in the future (see Section 5.5). The method used for the surveillance of the local mode S capable traffic (mode S is compulsory in terminal areas and airways) is that each aircraft broadcasts an unsolicited message every second at 1030 MHz called a *squitter*. This message includes the unique address of the sender and if it is received by another aircraft, the address is stored in the receiver's database of local traffic. Thus, every aircraft in the vicinity builds a list of all other traffic and can address each plane individually. The range of the other traffic is determined by the time lapse between interrogation and response and the bearing is derived from a directional antenna. Also, the responses from other mode S traffic will contain additional information including the altitude, heading, etc. The algorithm used to determine whether any of the local traffic poses a collision risk is described in Section 5.5 but having determined that a hazard is present the system will present to the pilot a "Traffic Advisory" (TA) or "Resolution Advisory" (RA). A TA alerts the pilot to be vigilant and prepare for evasive action if necessary while an RA requires immediate evasive action and the TCAS display will indicate what evasive action to take. Currently, this is limited to demanding a climb or descent and an indication of what rate is

required but future implementations of TCAS will incorporate left and right turns as well. The system operates independently of ATC and pilots are now trained to prioritize the information from the TCAS system following a collision over Überlingen in 2002 between two TCAS-equipped aircraft caused by one pilot following the RA and the other following advice from ATC.

1.6.3 Aircraft Communications Addressing and Reporting System (ACARS)

ACARS was first implemented in 1976 essentially as an automated reporting system for times spent in different phases of flight. The system accepts inputs from sensors on the wheels, doors, etc. to determine whether the flight is **O**ut of the gate, **O**ff the ground, **O**n the ground, or **I**n the gate (OOOI) and reports the data back to the airline operations. The communication is digital and uses whatever radio link is in use, that is VHF, HF, or satellite, as well as removing the burden for flight crews to record all this information the system provides for smooth operation, for example, by alerting replacement crews the optimum time to report for flight. The digital capabilities were soon expanded and the system can now be used to obtain additional information from airline operations such as weather reports at destination aerodromes, flight plans, amendments to flight plans, etc. It can also communicate with other ground facilities including the airline maintenance department and the aircraft or engine manufacturer. For example, ACARS can transmit a continuous report on the state of health of the engines to the manufacturer to organize maintenance schedules, troubleshoot problems, or provide data to implement design improvements. ACARS also communicates with ATC and is used to implement a new mode of communication between pilots and ATC known as Controller–Pilot Data Link Communications (CPDLC).

A major problem with voice communications, which were largely developed in an environment of one-on-one between pilot and controller, is that in a busy airspace a large number of pilots are on the same frequency. Any aircraft transmitting will block the frequency for all others and as the traffic increases, the probability that a pilot will accidentally cut off the transmission of another will also increase. In addition, each exchange between the controller and a pilot takes a specific amount of time and a traffic volume can be reached where it is no longer possible to pass the necessary messages to all aircraft within the time required. A solution is to increase the number of controllers and have each one on a separate frequency but that also introduces its own problems including the increased time required to handover flights from one controller to another. The CPDLC system alleviates these problems by using a datalink with ATC rather than voice communication to pass on clearances and other instructions to pilots and for pilots to request level changes, etc. This also relieves some of the stress on flight crews as they do not have to concentrate on picking out their

call sign from almost continuous ATC talk and the text information that appears on the screen cannot be misheard. ACARS controllers are also interfaced to a printer so that a hard copy of important information can be obtained.

As an addendum to this section, it is worth discussing the role that satellite communications and ACARS played in the attempts to find the missing Malaysian Airlines flight MH370 that disappeared on 7 March 2014. The system is a commercial service provided by Inmarsat that has to be paid for and Malaysian Airlines had opted only for the engine monitoring service that reports the state of engine health to Rolls Royce. The transponder was switched off and the ACARS system was disabled by switching off the SATCOM and VHF channels, which stopped all ACARS transmissions to the ground station. In this circumstance, the satellite continues to send a simple handshaking signal to the aircraft known as a *ping* every hour to check whether the aircraft is still online and the aircraft continues to respond. From the time taken between the ping and the reception of the aircraft response it is possible to determine the distance from the satellite. Six pings were sent after the loss of contact and from these it was possible to map out an arc traveled away from the geostationary satellite in two directions, North and South, but distance measurements alone could not distinguish which track was flown. Some innovative work by Inmarsat engineers analyzed the shift in frequency of the return signal due to the Doppler effect to determine the aircraft speed relative to the geostationary satellite. By comparing this with other Malaysian airline flights on the same route they were able to say with a reasonable degree of certainty that it was the Southern track that was taken, which was a significant help to the SAR operation. Tragically, the mystery of what happened to flight MH370 has never been solved but the incident has reignited the debate over whether full ACARS implementation should be compulsory to act as a kind of continuous "black box" in flight.

1.7 Development of Radio Navigation

1.7.1 Radio Direction Finding

In 1865, James Clerk Maxwell predicted the existence of electromagnetic waves, which travel at the speed of light and that light itself was such a wave. Between 1886 and 1889, Heinrich Hertz was the first to demonstrate conclusively that the waves generated by his spark gap transmitter were the same electromagnetic waves predicted by Maxwell, which was an important milestone in Physics. Almost immediately after the first detection of radio waves, the basic phenomenon that enables direction finding was also discovered by Hertz in 1888 using a simple loop of wire. He found that the maximum signal was detected when the plane of the loop was aligned along the direction of the transmitter and the signal went to zero when the loop was turned face on. The theory describing this

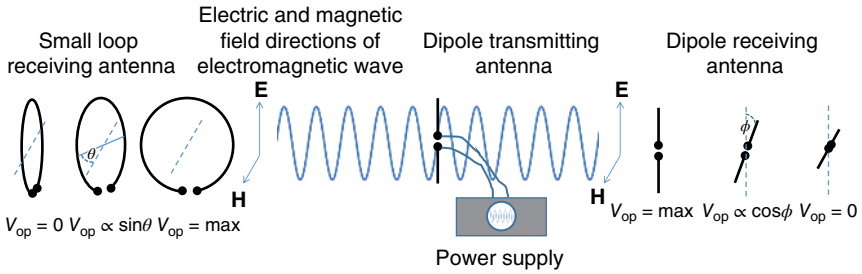


Figure 1.14 Schematic of a simple dipole antenna transmitting a linearly polarized wave. A small loop receiving antenna will give maximum signal when the loop axis is parallel to the magnetic field of the electromagnetic wave while a dipole receiving antenna will give maximum signal when it is aligned with the electric field.

observation is developed in Section 4.4.6, but the phenomenon is illustrated schematically in Figure 1.14.

Assume the radio wave is transmitted by a simple dipole aerial (not the case in Hertz's experiment but the characteristics of the radio wave are similar), which produces a polarized wave, that is, the electric field of the electromagnetic wave is vertical everywhere. This means that the magnetic field is aligned along the horizontal everywhere so that a loop set up as shown in Figure 1.14 is threaded by a time-varying magnetic field. If the loop diameter is small compared with the wavelength, the magnetic field can be assumed to be spatially constant. As shown in Section 4.4.6, in this scenario the time-varying magnetic field generates the maximum signal when it is aligned along the loop axis, that is, when the plane of the loop points to the transmitter. On the right-hand side of Figure 1.14 is shown the response of a dipole aerial and in this case the maximum signal is obtained when the transmitter and receiver dipoles are aligned and goes to zero when they are perpendicular. A simple loop does not give an absolute direction as it could be either side of the transmitter, that is, there is an ambiguity between two positions 180° apart. As shown in Section 7.1 this can be resolved by mixing the signals obtained from a loop antenna and a separate dipole antenna, also known as the sense antenna, to give a unique direction.

Small loops, that is, small compared to the wavelength, simplify the theory of coupling of the radio wave to the antenna but the signal is small and the use of small loops would have to wait till the development of electronic amplifiers. The reason that Hertz was able to detect the variation of signal with the loop orientation was that the transmitter was only a few meters away. Large loops that are resonant with the radio wave (that is, the total length of the loop is about half the wavelength) produce a much larger signal. They were already in use by the end of the nineteenth century and direction-finding patents were being filed by 1902. Given that it was the LF band being used with wavelengths of hundreds

of feet, these early devices were very unwieldy requiring massive loops. During experiments in 1907, two Italian engineers, Ettore Bellini and Alessandro Tosi, noticed that they could feed the signal from two large fixed perpendicular loops into two small perpendicular coils and recreate the directional properties of the radio wave in a small space. Thus, the movable coil could be placed in a bench-top device, the size of a saucepan, while the large receiving antennas could remain fixed. The patent for the Bellini–Tosi Direction Finder (BTDF) was filed in1909 and within three years the signal-amplifying abilities of the triode vacuum valve were first discovered allowing the BTDF to be combined with electronic amplification. These devices were in widespread use by the 1920s and continued in operation till the end of World War 2. Bellini and Tosi called their device a *radiogoniometer* and the basic principle, that is, recreating a large-scale electromagnetic field within a small desired space has found its way into several instruments in aviation including the ADF in use today (see Section 7.1).

Bellini–Tosi and other LF-based direction finders were successful instruments in widespread use for decades but the antenna arrays were much too large to install on an aircraft and they had to be ground-based systems with a ground operator reporting bearings to the pilot. One possible solution was to go to higher frequencies/shorter wavelength so that small loops could be used but this introduced a new problem. Moving in to the HF band, for example, where a 1 m loop would be resonant meant that radio waves would be reflected from the ionosphere and arrive at the aircraft from more than one direction. Thus, instead of picking up a single strong minimum in the signal, several weaker minima would occur making it impossible to determine the direction of the transmitter.

By the mid-1920s improvements in receiver and amplifier electronics meant that small multiple wire loops used in the MF band, just below frequencies where the ionosphere would become troublesome were sufficiently sensitive to use as direction finders and installed on aircraft. This meant that the ground station just had to provide an omnidirectional radio signal referred to as a Non-Directional Beacon (NDB). The earliest Radio Direction Finder (RDF) loops were turned manually or remotely by a motor with the operator finding the signal null and obtaining the bearing to the transmitting station (Figure 1.15a). By the end of World War 2 the loops would slew automatically to find the null if the signal was strong enough and pass the bearing indication to a cockpit display. The device was thus described as an ADF and the antenna was installed within a weatherproof aerodynamic housing as shown in Figure 1.15b.

The next development was a goniometer system in which the external rotating loop antenna was replaced by a pair of fixed loop antennae at 90° wound on ferrite cores to boost the field. This was then relocated via a goniometer system so a small rotating loop within the receiver could be used to find the direction of the NDB. In this design, there are no external moving parts on the airframe but since the 1980s it has been possible to implement the ADF entirely with solid-state electronics with no moving parts anywhere as described in Section 7.1.

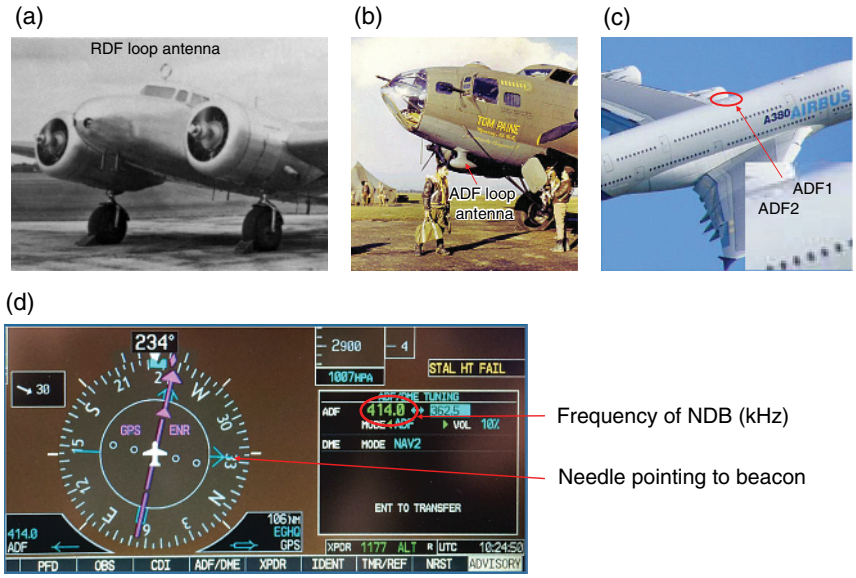


Figure 1.15 Evolution of Automatic Direction Finding (a) RDF rotatable loop antenna on Amelia Earhart's Lockheed Electra in 1937. *Source:* USAF. (b) ADF antenna in housing installed on a B17 at RAF Knettishall, England in 1943/44. *Source:* USAF Historical Research Agency. (c) ADF solid-state goniometers on the roof of an Airbus A380 in 2014. *Source:* Adapted from Ref. [8] and reprinted under Creative Commons license 2.0 [8]. (d) The ADF display in the PFD of a G1000 glass cockpit.

The compact ADF "blisters" for two independent ADF cockpit displays on the top of an Airbus A380 are shown in Figure 1.15c and the ADF display of the PFD in a glass cockpit is shown in Figure 1.15d. Generally, ADF used in conjunction with LF NDBs has limited use in radio navigation having been superceded by VHF navigation (see Section 1.7.3 and Section 7.2) and then Global Navigation Satellite Systems (GNSS: see Chapter 8).

Thus, it may be surprising to find a modern version of this radio navigation system that is over a century old implemented in the latest aircraft, but this is a testament to the power of an ADF and the intuitive information it provides. The needle simply points to the transmitter and shows the direction relative to the aircraft heading of the selected NDB. In modern displays, the outer angle scale of the ADF is slaved to the compass, as demonstrated in Figure 1.15d so that the bearing to the station can be read directly from the display. Such a device is referred to as a Radio Magnetic Indicator (RMI: see Section 7.1). Homing to a station is straightforward and many airport instrument approaches specify an approach procedure in terms of an NDB. In addition, NDBs are often used as anchor points for holds, for example, the holds North and South of

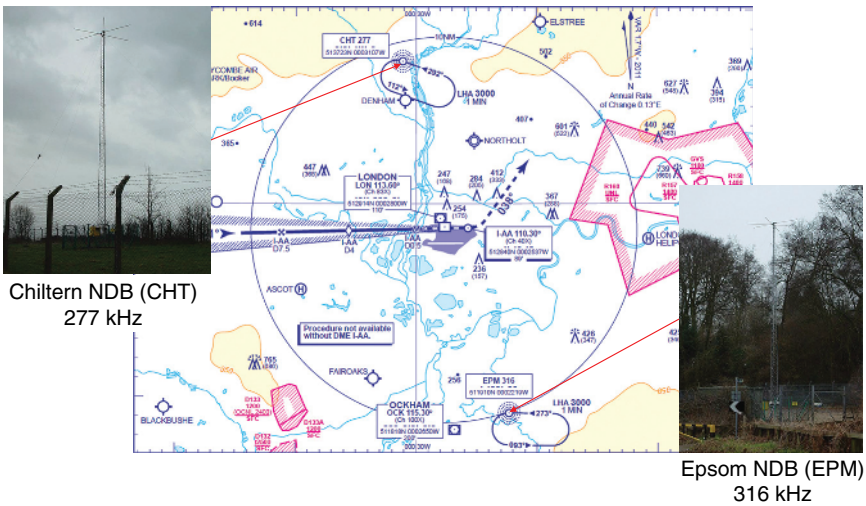


Figure 1.16 2015 approach chart for runway 09L at London Heathrow Airport showing holds North and South of the airport anchored on NDBs operating in the LF band.
Source: Chart reproduced with permission from National Air Traffic Services and photos of beacons reproduced with permission from Trevor Diamond (td@trevord.com).

London Heathrow Airport shown on a 2015 approach chart in Figure 1.16. These are anchored on the Chiltern NDB (CHT) and the Epsom NDB (EPS) both operating in the LF band. In all current NDBs the signal is modulated with a Morse code identification consisting of the three letters designating the beacon so that the flight crew can confirm that it is operating. In reality, the hold would normally be flown with the autopilot, which would use GPS to follow the racetrack pattern, with the position of the NDB stored in the database. The ADF system is there as a backup, however, and there are plenty of instances where it has been used due to failures of other equipment. Tracking to and from NDBs and flying holds using ADF is still an important part of training for commercial pilots.

1.7.2 Guided Radio Beam Navigation

Another form of radio navigation using the LF band is to set up a directional beam along which the aircraft flies. Systems started to be developed shortly after World War 1 based on a patent originally granted to the German engineer, O. Scheller, in 1904. The system envisaged by Scheller employs four overlapping beams as shown in Figure 1.17a. Each one is amplitude modulated to produce a string of Morse "A"s (• –) or "N"s (– •) as shown in Figure 1.17b and if these are picked up equally, that is, the aircraft is somewhere along the line of intersection of the beams, a continuous tone is heard in the headphones. Off the

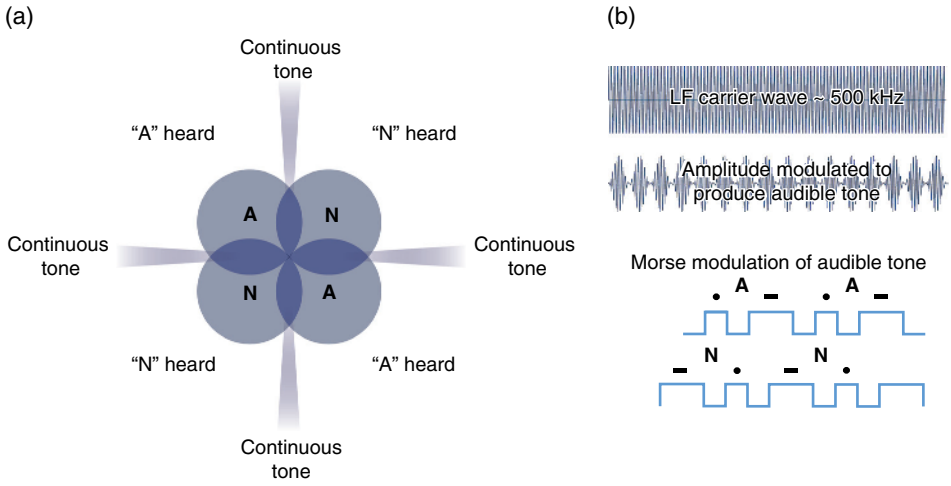


Figure 1.17 (a) Directional 4-beam system originally devised in 1904 by Scheller. A crossed loop or Adcock antenna generates the radiation pattern of overlapping radio beams. (b) Each quadrant is modulated by an audible tone, which itself is modulated by the Morse code for "A" or "N" in neighboring quadrants. If the two signals are picked up equally, that is, the aircraft is along one of the centerlines of the overlapping beams, then a continuous tone is heard. Moving off center produces an audible A or N indicates that the aircraft is off center and also which way to turn to recover the track.

intersection a Morse "A" or "N" is heard so not only does the system indicate whether the aircraft is off the centerline but it also shows which way to turn to recover it. The overlapping beam pattern was originally produced by crossed-loop antennas but subsequently a superior design of antenna invented by the British army officer, Frank Adcock, in 1919 was employed. This consisted of four vertical antennas connected underground with the current in opposing pairs in antiphase and was found to give a "cleaner" signal with reduced interference.

During the development of commercial air transport after World War 1, these beacons became widespread and in the United States an entire network was set up defining airways across the country. This was known as the Low Frequency Radio Range (LFR) and by the mid-1930s it covered the entire country with stations typically 200 miles apart. The transmitted beams defined airways that aircraft would navigate by tracking from one beacon to the next and a chart for the Silver Lake transmitter in California is shown in Figure 1.18. Thus, a pilot would be guided for 100 miles by the beacon behind and then switch frequency to the beacon 100 miles ahead. When flying over the beacon, the signal would disappear altogether momentarily then reappear on the other side indicating to the pilot they had flown directly overhead thus providing an accurate position fix.

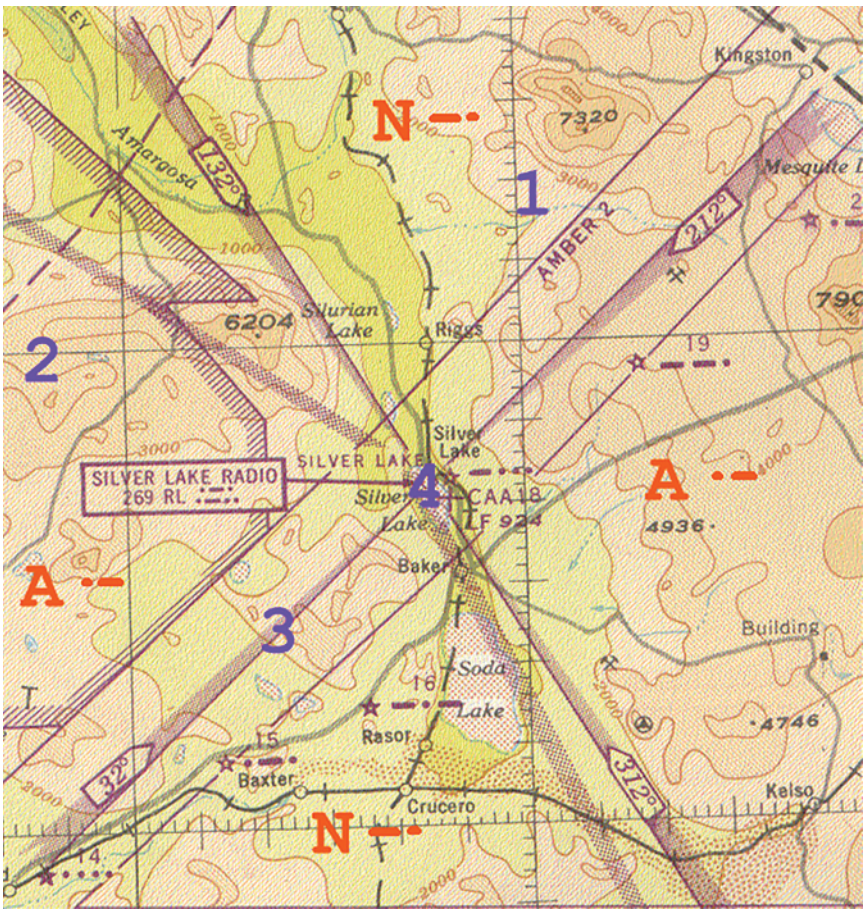


Figure 1.18 Silver Lake Low Frequency Radio Beacon defining four airways.

In addition to the Morse code "A"s and "N"s the stations would generate Morse identifiers every 30 seconds in the "N" quadrant and then in the "A" quadrant to confirm the identity of the station and indicate it was operating. From an operational point of view, listening to the continuous tones could be tiring and by 1930, cockpit instruments were available that would produce a visual "fly right" or "fly left" indication to maintain the centerline [9].

From the late 1920s, LFR networks facilitated the transition to reliable commercial air transport that, apart from extremes, could operate in all weathers and they dominated airways navigation for decades. By the end of World War 2, new VHF-based navigation systems had been developed (see next section) that addressed some of the problems with the LFR system. Being a LF

system made it susceptible to interference especially from electrical storms and the ground wave propagation mode of the LF waves could produce reflections and refractions off ground features giving false directional information. Another problem was that the beams from the beacons were restricted to just four directions. In principle, it is possible to produce six or eight beams from more overlapping radio signals but that creates its own problems. On the other hand, the later VHF stations could be used to fly along any chosen direction to or from the beacon.

The LFR network went some way to making navigation in air travel weatherproof but visual conditions were still required to land at the destination airport and from the late 1920s, short-range systems to guide aircraft into land were also being developed. In fact, the biplane that Doolittle performed his blind flight in 1929 was fitted with a prototype system using overlapping beams and a cockpit "fly left," "fly right" indicator to guide him along the runway center line. The instrumentation has many similarities to the Instrument Landing System (ILS) that is in widespread use today (see Section 1.7.3 and Section 7.4).

So, in principle, instrument landings were available from 1929 but Doolittles flight was something of a bold experiment and he also had a safety pilot in the back seat who was visual with the airfield and could take over if necessary. Landings in almost zero visibility (down to 75 m) can be performed by the autopilot in modern aircraft but only at airports that have the top level of ILS system (Category IIIB) installed and there is the requisite equipment on the aircraft including backup autopilots. There has to be some visibility on the ground otherwise the aircraft is unable to taxi off the runway. There is an even higher category of ILS defined (IIIC) with which a true zero visibility landing can be made, but this system also requires that the aircraft can taxi to the gate on the autopilot and at the time of writing this is not available anywhere in the world.

1.7.3 VHF/UHF Radio Navigation Systems

By the late 1930s, it was clear that future airways navigation beacons would need to operate at VHF frequencies to overcome the interference problems of the LF network and would also need to be able to provide flexible vectoring guidance to and from beacons. Since airways are defined by beacons that are 200 miles apart or less, the line of sight radio range needs to be 100 miles so, from Equation (1.1), the horizon is not a limitation for aircraft that are flying above 7000 ft, which is certainly the case for long-distance flights. The network went through a brief evolutionary phase using a VHF visual-aural radio range (see Interesting Diversion 7.1) but the development of an omnidirectional VHF beacon known as a VHF Omnidirectional Range (VOR) started in 1937. The first operational VOR transmitting at 125 MHz was put into service in 1946, and in 1949 the ICAO selected the VOR as the international civil navigation standard.

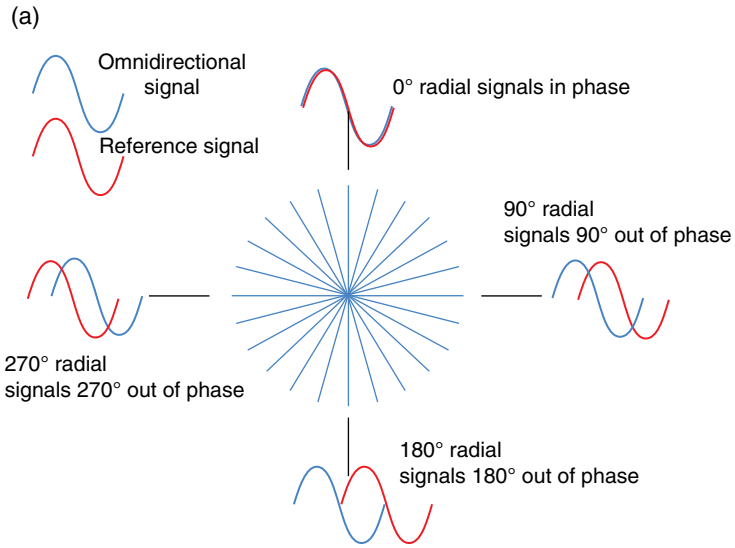
The operation of VORs is described in detail in Section 7.2, but in summary they determine the direction to or from the station by measuring the phase relationship of two 30 Hz signals modulated on a VHF carrier. The 30 Hz signals are an omnidirectional (reference) signal from a central antenna with the same phase along every azimuth and a directional (variable) signal, whose phase varies continuously around the circle from 0 to 360° relative to the reference signal (Figure 1.19a). The two signals are in phase along magnetic North, so comparing the phase angle between them gives the angle of the bearing to or from the station. A typical VOR transmitter is shown in Figure 1.19b and is installed near Daventry in the United Kingdom. The reference signal from the VOR is modulated to carry a Morse code three-letter identification (for example, Daventry is 'DTY', that is, – • • | – | – • – –) and is sometimes modulated by a voice channel to provide other information.

In the cockpit, there is a tuner to select the VOR frequencies and examples in a traditional and glass cockpit are shown in Figure 1.20a and b. In the latter case the VHF receiver can select two VORs (each driving its own separate display) and two in reserve whose frequency can be switched in by pressing a button. The VOR displays in the two cockpits are also shown in Figure 1.20a and b. Each has a knob known as the omnibearing selector (OBS) that rotates a pointer called the omni-bearing indicator (OBI), which shows the bearing selected to or from the station. This bearing represents the selected radial from the VOR and if the aircraft is on that radial the separate central part of the needle, known as the course deviation indicator (CDI) will be centered. Each division represents either 2° or 5° deviation depending on the type of display but in all cases full-scale deflection represents 10° off the selected radial.

There is also a TO/FROM flag to show whether the selected bearing is to or from the VOR. If the OBS is rotated through 360° the needle will center twice, one null will have the TO flag showing and the other the FROM flag showing. As an example, consider an aircraft on a heading of 045 approaching a VOR along the 225 radial (Figure 1.21a). Rotating the OBS will center the CDI when it is set to 045 with the TO flag showing and when it is set to 225 with the FROM flag showing.

An important aspect of the VOR display is that it is independent of the heading of the aircraft (unlike the ADF), it just gives an indication of where the aircraft is, assuming it is a point object. For example, in Figure 1.21b the three aircraft shown on the 225 radial will all display zero deflection on the CDI but only the aircraft with a heading of 045 will maintain zero deflection on the CDI. For the rest, the CDI will start to deviate as they move away from the 225 radial.

The VOR network achieved the design goals, that is, a VHF system relatively free of interference and flexible vectoring to or from beacons allowing a much greater flexibility of airways. The power of the system was increased further by installing at most VOR transmitters a separate UHF (300–3000 MHz)



(b)

Figure 1.19 (a) The VOR transmits an omnidirectional signal with the same phase in all directions and a directional signal whose phase relative to the omnidirectional signal varies from 0 to 360° around the circle. The two signals are in phase along magnetic North. (b) The DTY VOR in the United Kingdom. *Source:* Photo reproduced with permission from Trevor Diamond (http://www.trevord.com/navaids).

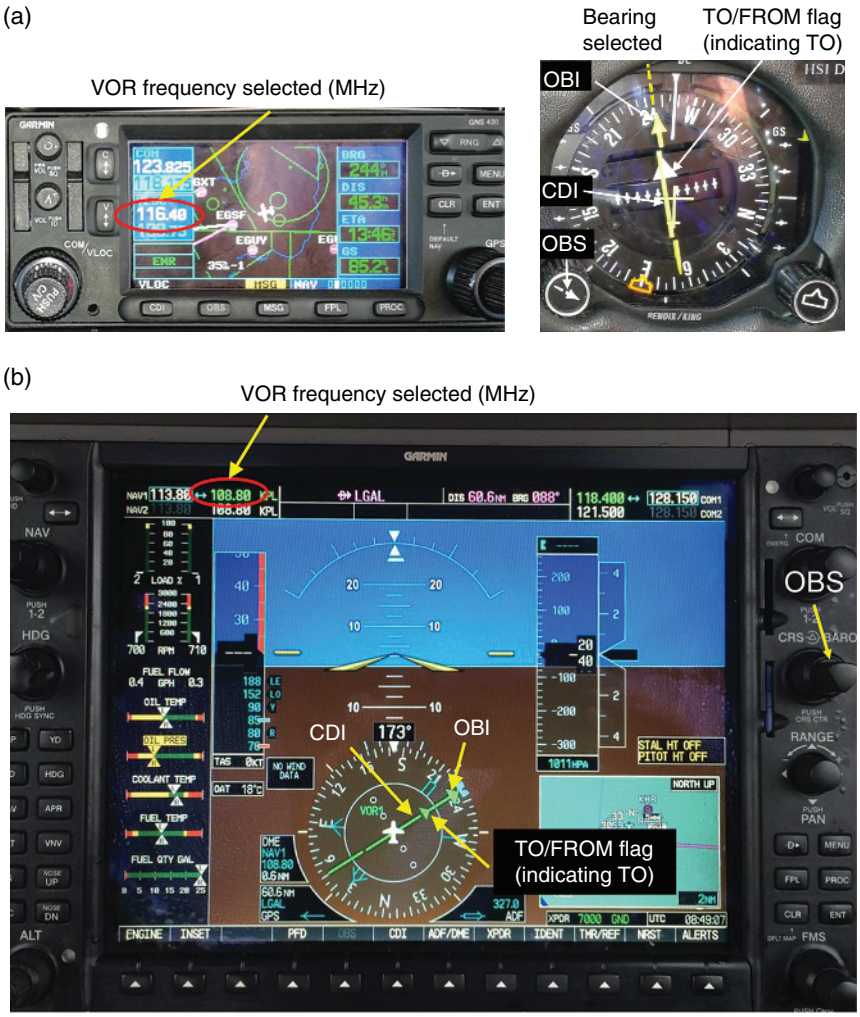


Figure 1.20 (a) The VHF Navigation tuner and VOR display in a traditional cockpit. (b) The VHF Navigation tuner and VOR display in a glass cockpit.

transponder that would transmit in response to interrogation by an aircraft UHF transmitter. The slant distance to the aircraft can be determined by measuring the time required for radio pulses to be returned. This Distance Measuring Equipment (DME) is described in detail in Section 7.3 and providing the pilot with a radial and a distance from the known position of the beacon enables an absolute position fix. In modern microprocessor controlled receivers the

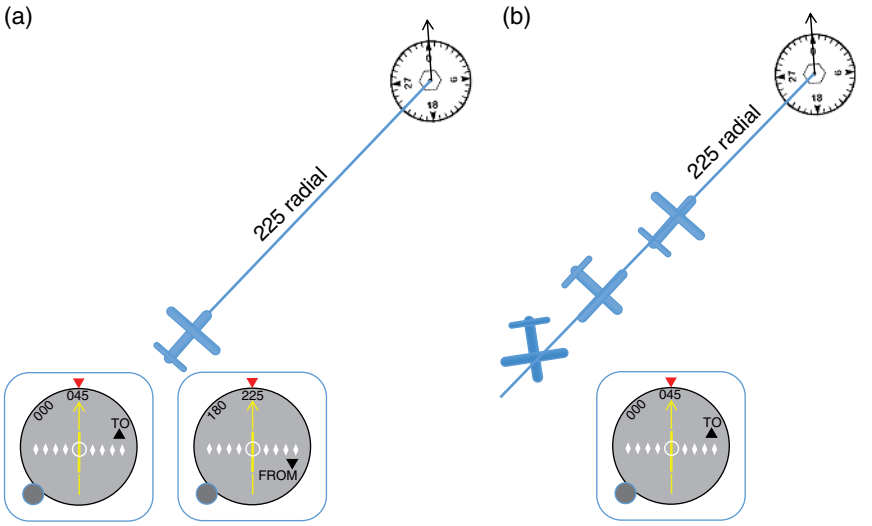


Figure 1.21 (a) An aircraft on the 225 radial will show zero deflection on the CDI for two OBS setting, that is, 045 with the TO flag showing and 225 with the FROM flag showing. (b) The indication of the display is independent of the aircraft heading.

pilot only needs to set the frequency of the VOR and the receiver automatically sets the UHF tuner to the corresponding DME frequency for that station from a database – a process known as frequency pairing.

With increased miniaturization of electronics allowing microprocessors to be incorporated into VHF navigation receivers, it became possible to increase the flexibility of the VOR system even further by allowing VORs to be electronically relocated within the receiver. For example, suppose the aircraft was navigating to a city airport with no VOR beacon but there was a beacon located 50 nautical miles to the West at a bearing 260 (Figure 1.22). The pilot could enter a shift of 50 miles along a radial 080 into the receiver, which would place a virtual beacon over the airport. The internal computer would obtain the radial and DME distance from the real beacon and compute what the phase shift and distance would be from the virtual beacon. The cockpit VOR and DME display would then be driven as if they were receiving signals from the virtual beacon. Thus, the pilot could use the instruments to navigate as normal as if there was a real beacon at the destination.

With this system, it is possible to plan any route between two points and not be restricted to flying from beacon to beacon. This increase in flexibility illustrates the important concept of area navigation (RNAV) to distinguish it from the linear point-to-point navigation that had been the norm since the 1920s. Restricting flights to airways does not normally provide the shortest route

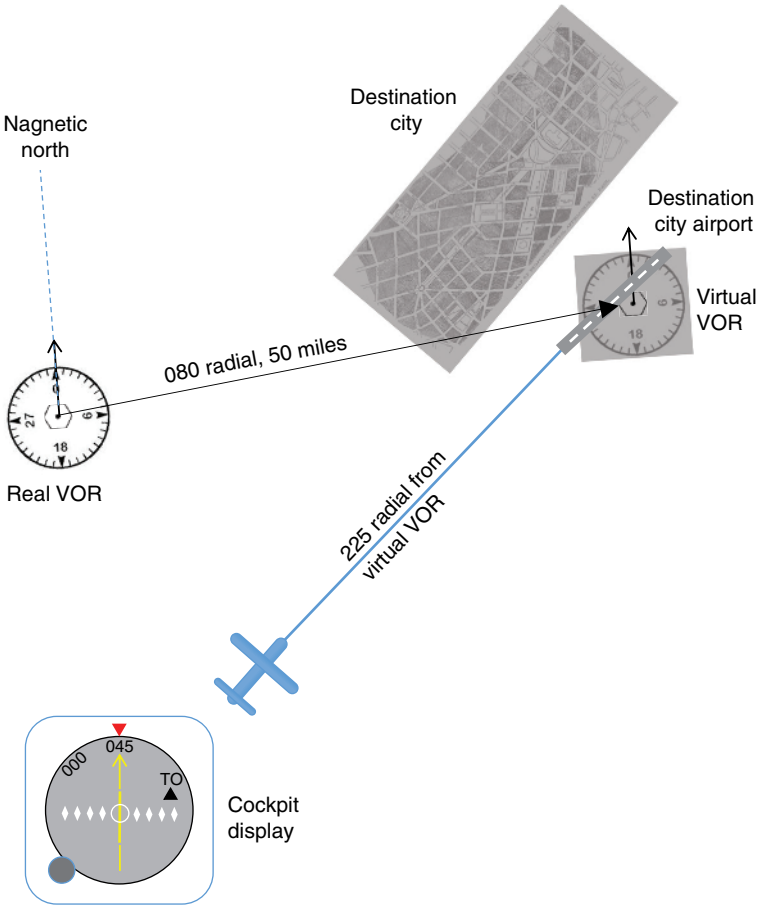


Figure 1.22 RNAV based on VORs. The pilot enters a radial and a distance to electronically shift a real beacon to generate a virtual beacon at the desired point. The cockpit VOR and DME displays are then driven as if they were receiving signals from a real beacon at the virtual beacon position.

between departure and destination points. With increased emphasis on saving fuel, it is rational to utilize the flexibility of the aeroplane to be able to go anywhere and fly routes that minimize the distance traveled, which would normally be great circle routes (see Section 6.1.2). The virtual VOR technology is only one of many RNAV systems, which today include GNSS and inertial navigation (see Section 1.8). Airways are still an important part of the traffic control system and flight plans are submitted using airways but wherever possible controllers will clear flights direct to destinations knowing that the necessary primary and backup RNAV navigation systems are on board.

The worldwide network of VOR beacons grew to over 3000 by the year 2000 with 1033 in the United States alone but this has started to reduce and at the time of writing there are 967 operating beacons in the United States. In 2015, 70 years since its inception, the VOR system has entered its twilight years with GNSS and inertial navigation becoming dominant in commercial aviation. Currently, the Federal Aviation Authority (FAA) in the United States is planning to reduce the network to 500 beacons by 2020 essentially to provide backup in the event of a GNSS outage and for GA aircraft not equipped with a satellite navigation receiver.

The discussion of VORs has focused on long-distance navigation but a VHF/UHF system was also developed for instrument landing. By 1929, the 4-course Low Frequency Range had already been used to provide radio guidance to line up with a runway. That year in the United States tests began on a HF system that could provide guidance for both the runway centerline and the required height at a given distance from the runway threshold, known as the *glideslope*. The system that was eventually developed is known as the ILS (see Section 7.4).

The ILS has the two components illustrated in Figure 1.23, that is, one to determine the deviation from the centerline (the *localizer*: Figure 1.23a) and one to determine the deviation from the required height (the *glideslope*: Figure 1.23b). The localizer antenna situated at the far end of the landing runway transmits two overlapping directional beams in the VHF band, one amplitude modulated at 150 Hz and the other at 90 Hz. The depth of modulation in both beams varies with angle away from the centerline (known as *space modulation*), so by measuring the difference in the depth of modulation (DDM) of the two frequencies the onboard instrumentation can determine the deviation from the centerline. This is passed to the same display as used for VOR tracking and when an ILS localizer frequency is selected the VHF receiver recognizes this and drives the CDI using the DDM measurement. In this mode, the turning of the OBS has no effect as there is only one direction (the runway heading) defined and also the sensitivity of the CDI is increased by a factor of 4 so that full-scale deflection corresponds to an angular deviation of 2.5°.

The glideslope works by a similar space modulation technique but the transmitted overlapping beams are in the UHF band and angled up at the glideslope angle, which is typically 3° relative to the ground. Within the receiver the UHF frequency is paired with the VHF frequency selected for the localizer, so it does not need to be set separately and the display depends on the type of cockpit. In a traditional cockpit, selecting the localizer frequency automatically switches in a second needle on the VOR display that is horizontal and whose vertical deflection shows the angular deviation from the glideslope with full-scale deflection corresponding to 0.7° (Figure 1.23c). In a glass cockpit selecting the localizer frequency automatically brings up an arrow display next to the height tape as

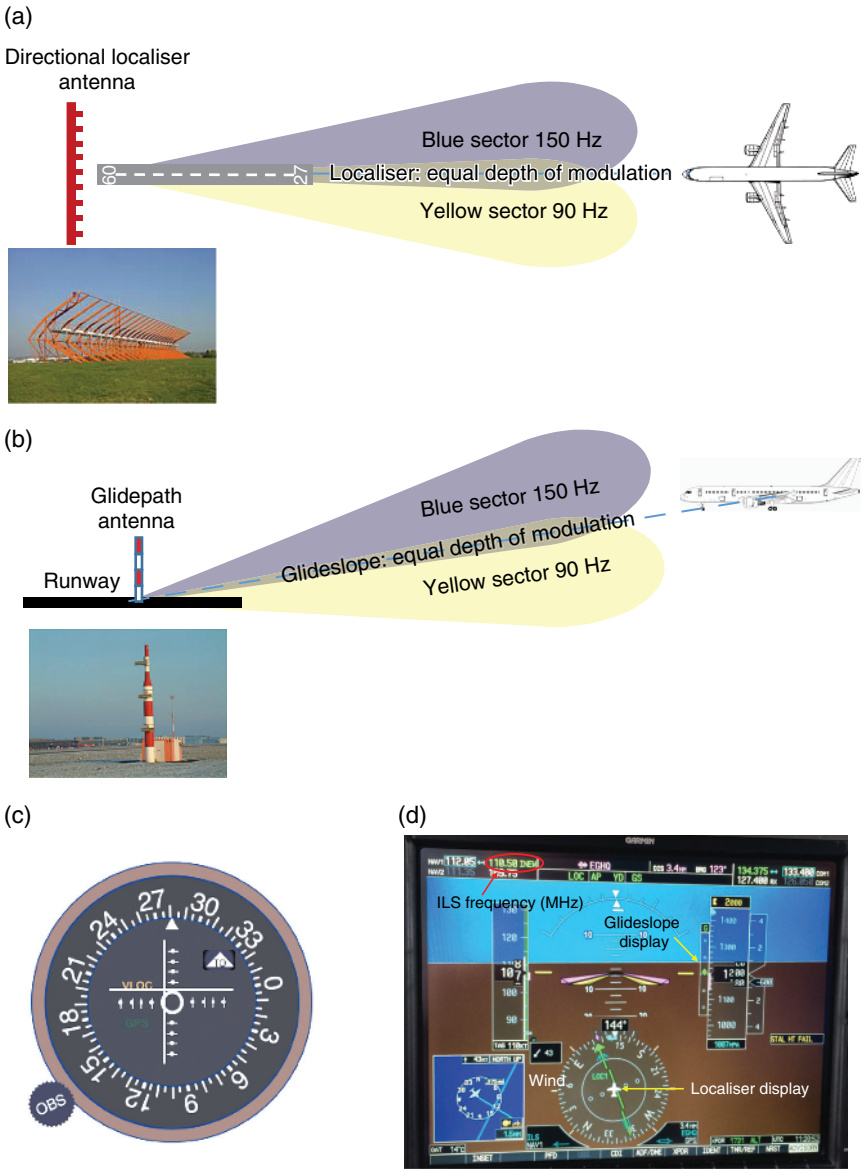


Figure 1.23 (a) Localizer antenna (typical example shown in photo) at the end of the runway transmits two overlapping directional beams amplitude modulated at 150 and 90 Hz with the depth of modulation (DDM) increasing away from the runway centerline. The difference in the DDM is used to determine the angular deviation from the centerline. (b) The glideslope antenna (typical example shown in photo) at the side of the runway produces similarly modulated directional beams and the DDM is used to determine the angular deviation from the glideslope. (c) Localizer and glideslope display in a traditional cockpit during an ILS landing. (d) Localizer and glideslope display in a glass cockpit during an ILS landing.

shown in Figure 1.23d, which also has a sensitivity of 0.7° deviation for full-scale deflection. Normally, the ILS localizer will be colocated with a DME transponder, described above and described in detail in Section 7.3, so that the distance from the runway threshold is known.

The system described is the one finally adopted internationally but in the United States, earlier systems that provided both localizer and glideslope guidance working at different frequencies were available before World War 2. The first landing of a commercial aircraft using ILS was in 1938 when a Pennsylvania Central Airlines Boeing 247D landed in Pittsburgh during a snowstorm. In 1941, the US Civil Aviation Authority authorized 6 ILS systems to be installed and by 1945, 9 systems were in operation with another 10 under installation. During the war, the US army had developed an improved system operating at higher frequency, which was adopted as the international standard by the ICAO in 1949 and is the system we have today. The signals that drive the displays can also act as control inputs to an autopilot and post-war there was a good deal of research effort dedicated to develop the technology to achieve a fully automatic landing without pilot input (see Section 1.9). This also needed the development of other technologies in addition to ILS, for example, the radio altimeter (see Section 5.6), but the first fully automatic landing was demonstrated in 1964 at the Royal Aircraft Establishment airfield at Bedford, England using a Hawker-Siddeley Trident jet airliner. The first fully automatic landing on a commercial flight with passengers was carried out at Heathrow airport in a British European Airways Trident in 1965.

ILS systems have evolved and are now classified into different categories depending on their authorized visibility limits and decision height, that is, the height below which a pilot should not descend if they are still not visual with the runway. The categories are dealt with in detail in Section 7.4.4 but as outlined in Section 1.7.2 above, the top category in operation (IIIB) allows descent to below 50 ft in visibility down to 75 m.

During the 1980s, a new system working in the SHF (3–30 GHz) called the Microwave Landing System (MLS) was developed that offered greater flexibility than ILS. For example, ILS is fixed at a given glideslope angle (usually 3°) along a fixed centerline, whereas MLS enables aircraft to choose a glideslope and approach at an angle to the runway. This is especially useful at military airports where a wide range of aircraft types varying from heavy transport to helicopters are operating. The system has only been implemented at very few civil airports and the number is not likely to increase as GNSS-guided approaches, which offer the same flexibility, are becoming available. ILS is still the dominant radio-guided landing system in use worldwide and its will remain operational for the foreseeable future. It is likely that over the next decade, GNSS-guided approaches will start to become the norm in operational practice but ILS will need to remain in place as a local backup.

1.8 Area and Global Navigation Systems

1.8.1 Hyperbolic Navigation

The concept of Area Navigation (RNAV) was introduced in the previous section and describes navigation systems that can be used to travel to any specific point in two dimensions as opposed to navigating along an airway. The ultimate aim has always been to achieve this globally so that an aircraft can be navigated to any point on the Earth without visual references. The earliest operational method was the British Gee system used by the RAF Bomber Command from 1942 and was based on a concept known as hyperbolic navigation. The basic idea was already well known in the 1930s but to build a practical system required equipment that could measure the timing of radio pulses at the microsecond scale and this came with the development of Radar in the late 1930s.

Hyperbolic navigation relies on the difference in timing between the reception of radio pulses from widely spaced beacons in chains with each chain consisting of a master and at least two secondaries. It is simpler to describe the method with a specific example and Figure 1.24 shows a master and one secondary separated by a distance of 100 nm, which is known as the baseline. The master transmitter emits a series of radio pulses typically at a rate of 10 a second and Figure 1.24a shows the timing of one pulse as perceived by an aircraft, labeled A, at the midpoint on the baseline given that the waves propagate at 6.18 μs per nautical mile. At $t = 0$ the pulse is emitted by the master and at $t = 309$ μs it arrives at the aircraft and is registered on a display. At $t = 618$ μs the pulse arrives at the secondary, which then transmits its own pulse after a precise time delay – 100 μs in this example, that is at $t = 718$ μs. This pulse arrives at the aircraft at $t = 1027$ μs and is registered on the display, which thus shows a time difference between the pulses (TD) of 718 μs. Now consider another aircraft, labeled B, on the same bisecting line. It will take a longer time, say t_1, for the direct pulse from the master station to reach it. The secondary will still emit its pulse at $t = 718$ μs, which will also take time t_1 to reach the aircraft, so when the difference is taken the two t_1's will always cancel. It is clear that anywhere along the bisecting line the time difference is always 718 μs, that is, the time for a pulse to traverse the baseline plus the delay introduced by the secondary.

It is intuitive that any other constant TD value will also be measured along a line, which will be curved as we move away from the bisecting line. It can be shown that the curves are hyperbolae as illustrated in Figure 1.24b and this is the origin of the term hyperbolic navigation. Measuring the TD value will then locate the aircraft on a specific hyperbolic line and to obtain a position fix, another secondary with its own set of TD hyperbolae is used as shown in Figure 1.25a. The pulses from the three stations are encoded in some way (for example, different numbers of closely spaced multiple pulses) so that they

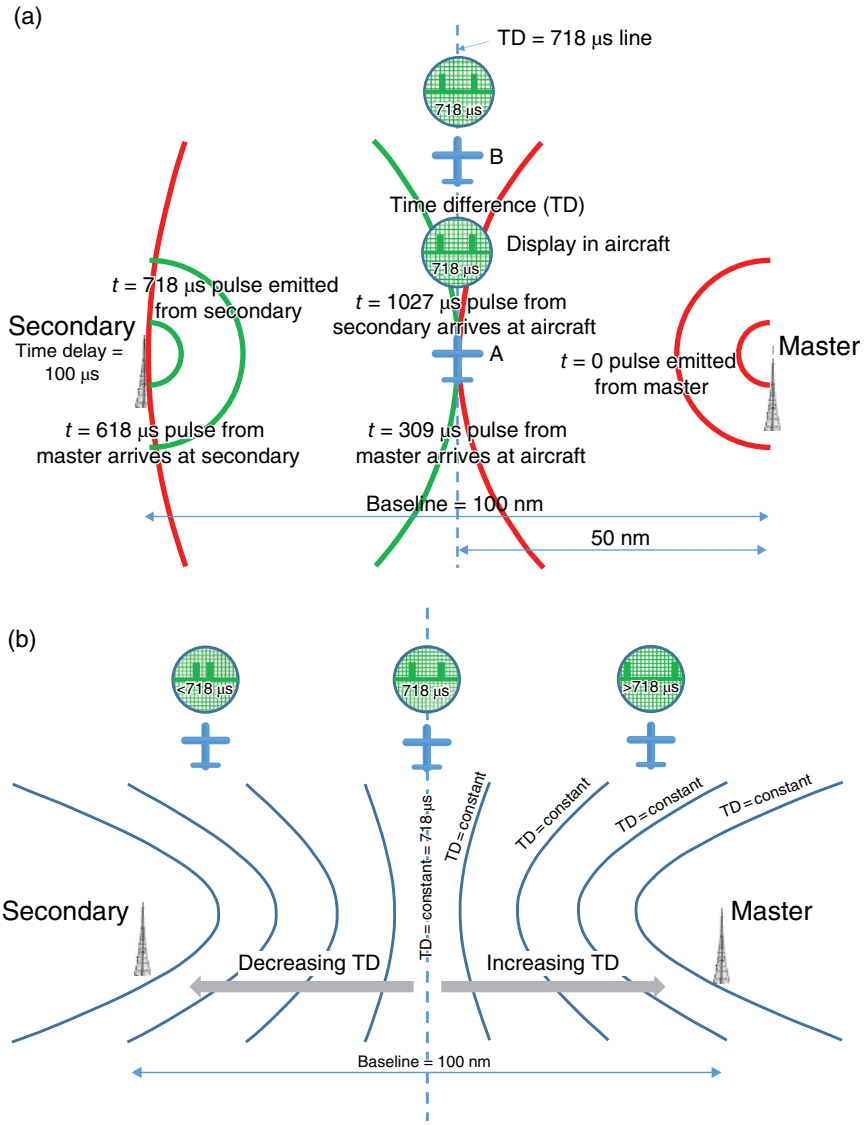


Figure 1.24 (a) Pulse timing for a specific example of a master and secondary station 100 nm apart and a secondary time delay of 100 μs. Anywhere along the bisecting line between the two stations an aircraft will measure a time difference (TD) of 718 μs. (b) Away from the center line the lines of constant TD form hyperbolae.

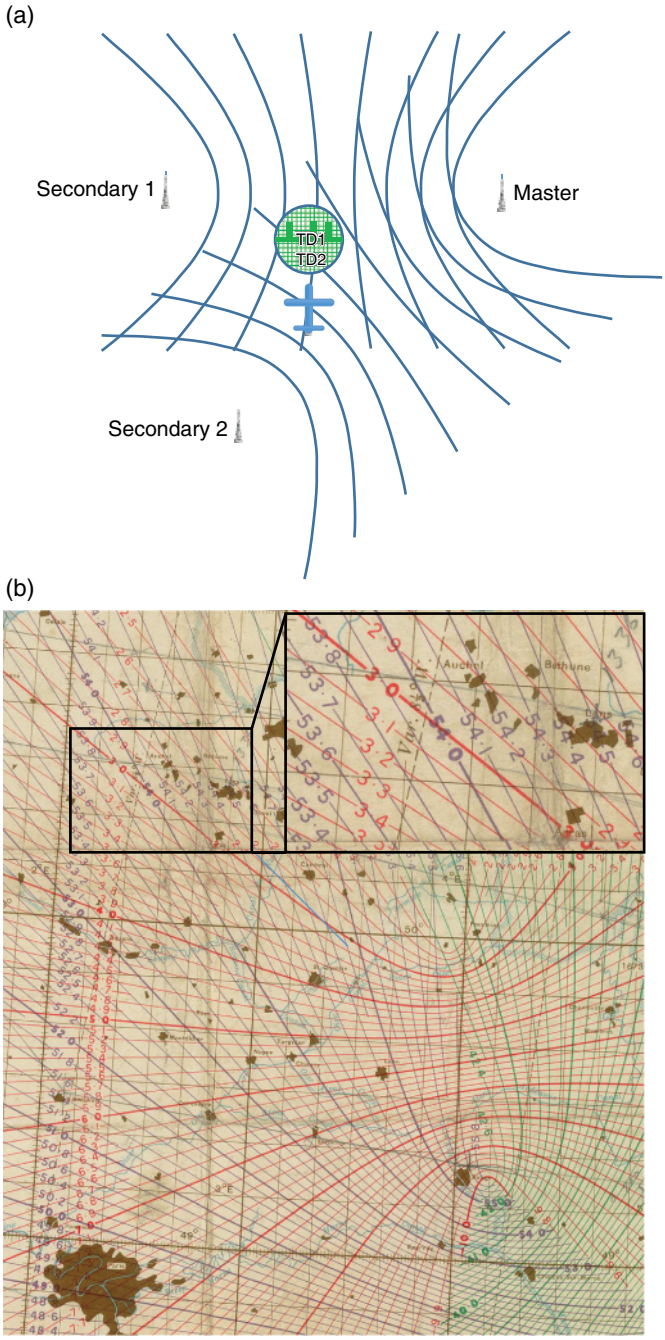


Figure 1.25 (a) To get a fix at a single point a second overlapping set of constant TD hyperbolae are required from another secondary in the chain. (b) A Gee navigation chart for the Reims chain operating in 1944 with the inset detailing the color-coded TD values. *Source:* Reproduced with permission of Smithsonian Institution office of imaging, printing and photographic services. [10].

can be distinguished. Notice that the entire system is synchronized by just the master station and there does not need to be any physical connection between different transmitters in a chain. During the war, the TDs were measured in the aircraft using an oscilloscope trace and having obtained them the values were plotted on a chart of constant TD lines for a particular chain. An example is the chart for the Reims chain of the Gee system in 1944 shown in Figure 1.25b with the inset detailing the color-coded TDs. The Gee system had an accuracy of a few hundred meters at a range of 350 miles and remarkably, was in operational use by the RAF till 1970.

The British Gee system was operational by 1942 and as such the first genuine RNAV system but it was relatively short range and after the invasion of France in 1944, new chains were set up as the front progressed. It operated at relatively high frequencies around 30 MHz, that is, at the top of the HF band so that sky-wave contamination could be a problem. The US was close behind and by 1943 the US coastguard had a hyperbolic navigation system operating that was known as LORAN (LOng RAnge Navigation). This system, whose designation became LORAN-A to distinguish it from later systems, expanded rapidly and by 1945 included over 70 transmitters that provided navigation over 30% of the Earth's surface. By the late 1940s, experiments had shown that reducing the carrier wave frequencies into the 90–110 kHz (LF) band produced significant improvements in range. This change and other developments resulted in a new system LORAN-C, which by the late 1950s became the global all-weather radio navigation standard for ships and aeroplanes. LORAN-A and LORAN-C were both operated in parallel till the mid-1970s when LORAN-A started to be phased out.

The absolute accuracy (systematic error) of the LORAN-C system was of the order of 200 m while the repeatability (random error) could be as low as 20 m with the onboard receivers automatically measuring TDs and displaying latitude and longitude. In the 1960s, the required instrumentation was expensive and used primarily by the military but following the reduction in cost of solid-state electronics in the 1970s, LORAN-C found its way into widespread civil use and even into the GA market as shown by the installation in Figure 1.26.

After GNSS became widely available in the late 1990s, LORAN-C started to fall into disuse and the decision was taken by the US Government to terminate the signals in February 2010. Shortly afterward the LORAN-C operated by the Canadian government was shut down and a similar hyperbolic navigation system run by the Russian government (CHAYKA) was also terminated. This did not mark the end of hyperbolic navigation systems, however, as increasing fears about the vulnerability of the GNSS system to natural and hostile interference has prompted plans to provide a backup for GNSS and this has led to the emergence of eLORAN (for enhanced LORAN). This is envisaged to reach an accuracy of ±8 m, which is competitive with un-enhanced GNSS. The first chain of

Figure 1.26 LORAN-C installed in a General Aviation aircraft (Grumman AA5) in the 1990s alongside a GNSS display. *Source:* Reproduced with permission of Rob Logan.

eLORAN transmitters became operational in the United Kingdom in October 2014. This technology, however, is emerging at the same time as tests have begun on a cheaper terrestrial navigation system based on tracking transponder signals (see Section 5.4.12).

1.8.2 Global Navigation Satellite Systems (GNSS)

GNSS are the youngest of the global navigation technologies as they were not conceivable before the dawn of the space age starting with the launch of the Sputnik capsule in 1957. Satellite-based navigation systems started to appear shortly after, however, and the first navigation satellite, which formed part of a US Navy system called Transit, was launched in 1959. The complete system consisted of seven satellites in a low polar orbit (see Section 8.2 for a description of different types of orbit) that transmitted stable radio signals, which were used by surface craft to determine their position by measuring the Doppler shift of the radio waves. There was also a series of ground stations to track the satellites and update their orbital parameters for the users. Although Transit worked on a different principle to GNSS, it tested various elements of the infrastructure required for future GNSS. The system was made available for civilian use in 1979, becoming widespread among commercial ships and it continued in operation till the late 1990s. Some drawbacks of the Transit system, including its inability to determine height, the long monitoring times required to obtain a position, and the relatively long off-air periods made it unsuitable for air transport.

A second system developed for the US Navy, called Timation, tested the application of high-stability clocks, the accuracy of time transfer, its ability to determine position, and the initial tests of three-dimensional navigation. The first satellite was launched in 1967 in which the timing was provided by stabilized quartz clocks but later missions flew the first-ever atomic clocks into orbit, which were sufficiently precise and stable to provide accurate position fixes. Thus, the fundamental concept by which GNSS determines the position was developed and tested with the Timation system.

The principle of the position fix is illustrated in Figure 1.27. Each satellite transmits its current orbital position (ephemeris) and since it carries a very accurate atomic clock, it knows precisely when the message is sent. The broadcast is picked up by the vehicle and imagine for a moment that there was also an atomic clock in the receiver synchronized to that of the satellite. Then, knowing the position of a satellite (contained in the message) and the time it was sent, the time for the signal to reach the vehicle and thus the distance to the satellite would be known. A single satellite would define the position of the vehicle as somewhere on the surface of a sphere (Figure 1.27a). Two satellites would fix

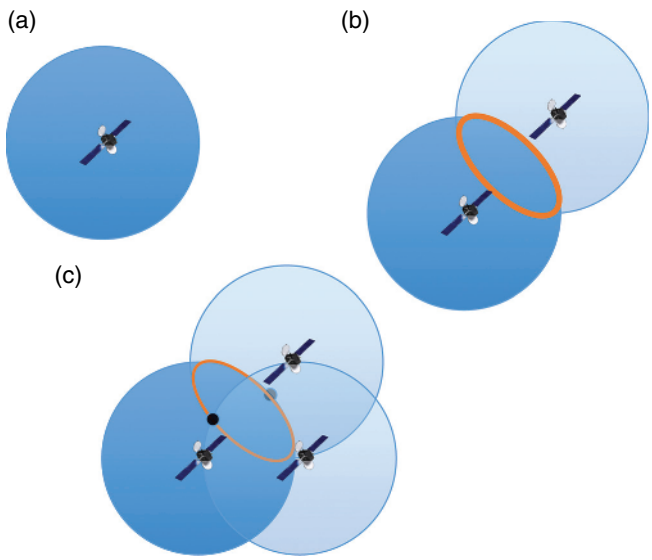


Figure 1.27 Principle of obtaining a position fix from satellites. If there was an accurate clock on the receiver synchronized to that of the satellite, then the distance to each satellite would be known independently so (a) one satellite would define a point on the surface of a sphere, (b) two satellites would define a point on a circle, and (c) three satellites would define two points, one of which was far out in space and could be discarded. In practice, there is not sufficiently accurate timing available in the receiver so a fourth satellite is required to solve equations for *x*, *y*, *z*, and time.

the position as somewhere on a circle (Figure 1.27b) and three satellites would place the vehicle at one of two points (Figure 1.27c), one of which will be near the surface of the Earth and the other will be far out in space and can be discarded. The problem is that we do not have a sufficiently accurate clock on the receiver so that a fourth satellite is required to solve four simultaneous equations for x, y, z, and time, thus a by-product of getting a position fix is that we also know the time very accurately. If signals can be obtained from more than four satellites, the accuracy of the position fix is improved. The x, y, z coordinates determined by the system define a point in space but to relate this to a geometrical position on the Earth's surface in terms of latitude and longitude, the system needs to know the exact shape of the planet's surface (see Section 6.4).

Simultaneously, the US Air Force was working on its own GNSS and by 1972 had demonstrated the Pseudo-Random Noise (PRN) format of the satellite radio signals that would become the norm for GNSS (see Section 8.3). In 1973, a new program, the Navstar global positioning system (later to become just GPS), emerged by combining the Navy and Air Force systems. Between 1975 and 1985, 11 satellites designated as Block I were launched to begin a program of testing of the system and simultaneously a ground program to test user equipment began. Some of the Block I satellites also carried sensors to detect nuclear detonations partly as a contribution to monitoring of the test ban treaty.

Despite some early wobbles with funding and also, following the Space Shuttle Challenger disaster, a decision to switch to conventional rockets for launch vehicles, 24 satellites designated as Block II and (after further improvements) Block IIA were launched starting in 1989. This constituted the first fully global *constellation* of space vehicles. Only two satellites of this original constellation are still operational having been replaced with new vehicles with additional signaling capabilities designated as Block IIR, Block IIR(M), and Block IIF (Figure 1.28).

From 1976 onward, Russia developed an independent GNSS knows as GLONASS (GLObal NAvigation Satellite System) and began launching space vehicles in 1982. The full constellation was in orbit by 1995 but then went in to a partial decline due to a lack of funding. From 2001, the system became a spending priority and was fully restored by 2011 to provide global coverage. The orbital parameters for GLONASS are slightly different to those of GPS but the signal formats are similar.

The GPS system was made available for civil use at no charge in 1993, though the clock signal was deliberately dithered to limit the accuracy of position fixing to 100 m for civilian users – a process known as selective availability. In 1994, the FAA approved GPS as a stand-alone aircraft navigation system for all phases of flight and non-precision approaches. In 2000, US president Bill Clinton announced the removal of selective availability and overnight the precision of GPS position determination for all users improved to around 10 m.

Figure 1.28 Final GPS IIR(M) satellite being launched by a Delta II rocket in 2009. *Source:* Courtesy of gps.gov.

GLONASS and GPS are available to civilian users on a similar basis, but both satellite constellations transmit additional messages that are only available to the militaries of the respective countries. Modern GNSS receivers use both GPS and GLONASS signals to improve the satellite coverage and the precision of position determination. GPS/GLONASS without further improvement is still not accurate enough for precision approaches, but in recent years a number of augmentation systems, effectively add-ons to GNSS (described in Section 8.6), have been implemented to increase the accuracy of position determination and GNSS precision approaches are now available at 1746 airports in the United States alone [11]. Such procedures are known as Localizer Performance with Vertical Guidance (LPV) approaches and are also designated as RNAV procedures in recognition of the fact that Flight Management Systems (FMSs) on modern aircraft are normally using more than one navigation system.

A decision was made by the European Union to develop an independent GNSS amid concerns that the GPS and GLONASS systems could be disabled for non-US/Russian users at any time and an agreed program among participating nations started in 2003. This system is designed from the outset for civilian use unlike the GPS and GLONASS systems that were developed primarily for the military. The system will have its own built-in space-based augmentation system provided by additional geostationary satellites and will also combine

position fixing with additional SAR services. Each satellite is equipped with a transponder that will be triggered by a user's emergency locator transmitter (ELT). Signals will be sent to the relevant rescue co-ordination center (RCC) with a precise location of the user in distress. The service will be free to all users at the basic precision level and at a charge for the high-precision augmented service, which has an accuracy of around 1 m. The first satellite launch was in 2011 and the system was declared operational at the end of 2016 but the full constellation will not be complete till mid-2018 so there are still holes in the coverage. Other systems around the world include the fully global Chinese COMPASS/BEIDOU system due to be operational by 2020 and partial coverage systems operated by Japan and India. All of these GNSS systems require a significant ground infrastructure of tracking and communication centers, but these are covered along with the detailed operation of GNSS in Chapter 8.

1.8.3 Inertial Navigation Systems (INS)

Inertial navigation is unique among the systems available in that it determines the position entirely by measuring acceleration within the vehicle itself requiring, in principle, no external input whatsoever. The only restriction is that the starting position needs to be known but then all subsequent positions during travel can be deduced. The basic principle is that if acceleration can be accurately measured along an axis, then integrating the acceleration with respect to time gives the velocity along that axis and integrating again with respect to time determines the distance traveled as illustrated in Figure 1.29. This shows an accelerometer measuring motion in one dimension over a period of 10 minutes and initially there is an acceleration of 0.1 g, which produces an increase in velocity to 111 m s^{-1}, followed by a period of no acceleration (constant velocity). Then there is a deceleration of –0.05 g, which reduces the velocity to 58.5 m s^{-1} followed by another period of no acceleration and constant velocity. Finally, there is a further deceleration that brings the object to rest. During periods of acceleration (deceleration) the distance increases at an increasing (decreasing) rate and during periods of no acceleration (constant velocity) it increases at a constant rate. Finally, when the object has been brought to rest the distance stops increasing and the object has traveled 36.6 km, which has been determined solely by measurements with the accelerometer.

If we now put two accelerometers, one aligned along North–South and the other along East–West on a platform, and we know the starting position it is possible to determine the distance traveled along N–S and E–W independently and hence track the position on the Earth. A problem is that unless the platform is precisely perpendicular to the force of gravity, this will be measured as an acceleration and produce erroneous calculation of position. In the original inertial navigation systems, the platform on which the accelerators are mounted was kept perpendicular to Earth-vertical by mounting it in a gyroscope-controlled

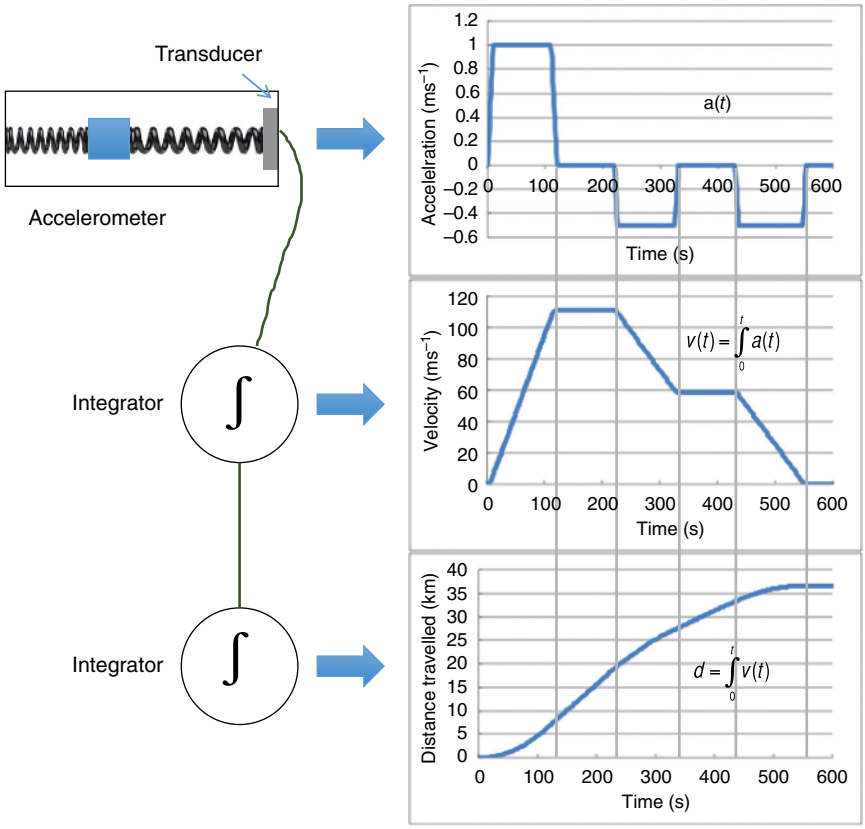


Figure 1.29 Principle of inertial navigation in one dimension. Integrating the measured acceleration gives velocity and integrating the velocity gives the distance traveled.

gimbal system that would maintain orientation while the aircraft pitched and rolled as illustrated in Figure 1.6. There are other complicating factors such as the aircraft traveling over a curved rotating Earth, which means that the platform has to be constantly adjusted to maintain itself perpendicular to gravity as described in detail in Chapter 9. In more modern systems the platform is fixed immovably to the aircraft and gyroscopes are used to determine the flight attitude and thus the component of Earth's gravity being measured so it can be deducted from the integration in software. These so-called *strap-down* systems are also described in detail in Chapter 9 and in this section the focus will be on the historical development.

The concept of inertial navigation was around before World War 2 but the first operational inertial guidance systems, albeit a partial version, was used in the German V2 (or A4) rockets in 1944. In the guidance system beneath

the warhead was an accelerometer with a single integration stage that could determine velocity. The engineers would calculate what velocity was required to be reached after taking off from the starting position in order to reach London (and later other cities) ballistically. The system would then be set to shut off the engine at the required velocity and the missile would continue on a ballistic trajectory reaching its target if the calculations were correct. The system was crude and had an accuracy in the region of 5 km but this was later improved by using additional radio guidance.

Post war, the focus of development in INS was for use in missiles and the cold war spurred the design of smaller, lighter, and more accurate systems. They also played an important role in space exploration as probes that are sent into interplanetary space have no other navigation system to rely on apart from tracking by Earth stations. It was the military that initiated the use of INS in aircraft and the first flight test was in 1958. By the early 1960s many strike aircraft had an INS and a typical example is the Litton LN-3-2A system shown in Figure 1.30. These early platform-stabilized INSs were very expensive precision-engineered assemblies that could shield the accelerometers from Earth's gravity within an aircraft that could produce accelerations of between −5 g and +9 g in maneuvers. In order to attain the required accuracy, the gimbal assembly,

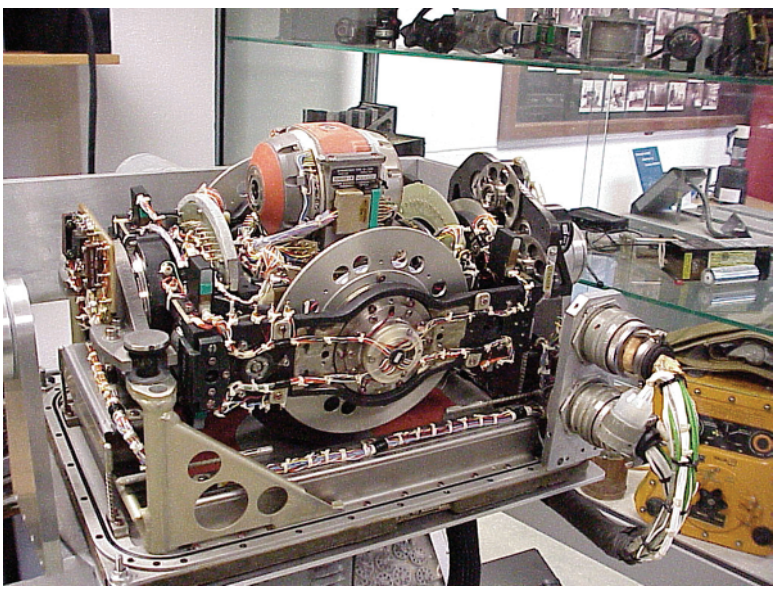


Figure 1.30 Litton LN-3-2A Platform-stabilized INS used in military strike aircraft in the early 1960s. *Source:* Reproduced with permission of https://en.wikipedia.org/wiki/LN-3InertialNavigationSystem. Licensed under CC BY-SA 3.0 [5].

accelerometers, and electronics all had to be kept in a temperature-controlled environment. In any INS system, the accuracy degrades with time and for the LN-3 devices this was quoted at 2 nautical miles error per hour of running.

The early systems were too expensive to use in commercial aviation but costs gradually reduced and in the late 1960s Delco introduced the popular Carousel INS system, which was installed on 707, 727, 737, 747, DC10, and Tristar airliners. This was also a gyro-stabilized platform INS and is shown in Figure 1.31a with the Earth-centered platform containing the accelerometers indicated. The control units in the cockpit were the Control Display Unit (CDU) and the Mode Selector Unit (MSU) (Figure 1.31b). When the INS was first switched on (MSU switched to STBY), the current position of the aircraft would be entered on the CDU and the MSU would then be switched to ALIGN. This would initiate a process taking approximately 10 minutes in which the gyro-stabilized platform set its attitude to obtain zero measured acceleration, during which the aircraft had to be at rest. Other procedures were also carried out during this initialization process to obtain the direction of true North and are described in Section 9.2.4. Waypoints would be entered into the CDU and when the system was ready it could be switched to NAV from which point any movements were recorded by the accelerometers. After takeoff, the unit could drive the autopilot to fly the aircraft to its destination via the entered waypoints. The availability of INS in aircraft from the late 1960s produced a step change in the ease of long-distance navigation, especially over oceans where radio navigation aids

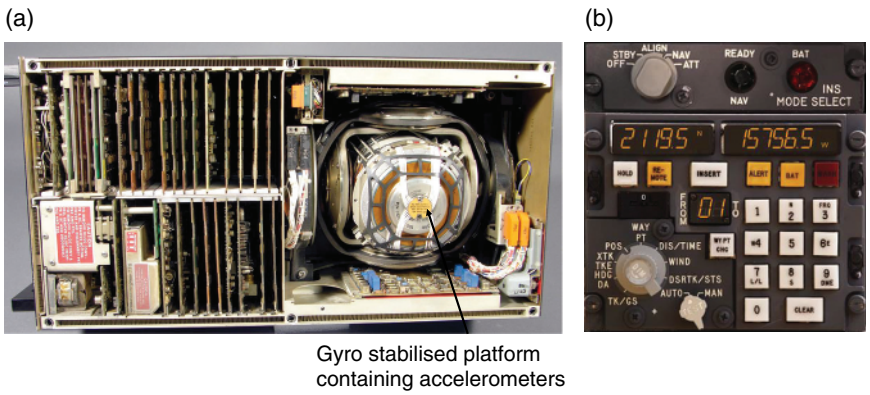


Figure 1.31 (a) Hardware and control electronics of the Delco Carousel with the case open to reveal the gyro-stabilized Earth-centered platform containing the accelerometers. *Source:* Reproduced with permission of National Air and Space Museum. (b) Lower part: Control Display Unit (CDU) used to enter the position at startup and waypoints. Upper part: Mode Selector Unit (MSU) to turn the unit on, initiate the align procedure, and select Navigation mode. *Source:* Screenshot of the CIVA simulator for X-Plane reproduced with permission from Philipp Münzel.

were not available. It remained the main tool of long-distance navigation till GNSS was available in the mid-1990s and thereafter was always used in combination with GNSS. Prior to INS, LORAN-C was available but did not have complete global coverage and was much more awkward to use than INS. A significant aspect of oceanic navigation prior to INS had been a combination of dead reckoning (see Section 6.5) and star sightings with some aircraft being fitted with periscopic sextants.

Despite mass production, INSs with gyro-stabilized platforms were still very costly due to the highly demanding engineering required to provide an Earth-centered platform for the accelerometers of sufficient stability to maintain acceptable navigation errors. This prompted an alternative approach in which the accelerometer platform is attached directly to the frame of the aircraft and gyroscopes mounted on the platform measure the pitch, roll, and yaw. If the attitude of the aircraft is known, a computer can then transform the measured output from the accelerometers to what the values would be if the platform was exactly perpendicular to gravity. This *strapdown* approach relies to a large extent on the availability of fast powerful processing since measurements from all accelerometers and gyroscopes have to be taken at 1000s of times a second and processed at that rate. Airborne computers in the 1960s built from discrete transistors were not up to the job but the microelectronic revolution starting in the late 1960s eventually made sufficiently powerful processors available.

Another technical requirement that needed to be met for the strapdown systems to work was sufficiently accurate gyro measurements of the attitude of the INS platform. The preferred option is to measure the rate of rotation of the platform by *rate gyros* (see Section 3.1.10) and from this compute the final angle relative to a given aircraft axis that the platform has rotated to. In the 1960s, strapdown rate gyros were not sufficiently accurate to achieve acceptable navigation performance. An alternative to rotating mechanical gyros is an optical method using a phenomenon known as the Sagnac effect, discovered in 1913 by the French physicist Georges Sagnac. He showed that when rays of light move in opposite directions around a circular cavity on a turntable, the light traveling with the rotation arrives at a target slightly after the light traveling against the rotation. The discovery of lasers in the 1950s made it possible to observe tiny changes in the interference pattern of two beams in circular paths and a practical Ring Laser Gyro (RLG) (Figure 1.32: see Section 3.2.4) first appeared in 1963 [12]. Since changes in interference patterns can be measured with high sensitivity, the RLG can detect small rotations of the table on which it is located and was a good candidate to act as rate gyro. Strapdown INS systems using RLGs were first tested in missiles in 1974, then in commercial aircraft in 1978, and were first used in passenger flights on Boeing 757 and 767 airliners [13]. From the start, at least two independent INSs were installed as redundant units so that if one developed a fault, navigation could continue with the other.

Further improvements came in the early twenty-first century with the development of MEMS (see Section 3.2.6) accelerometers used on the INS platform.

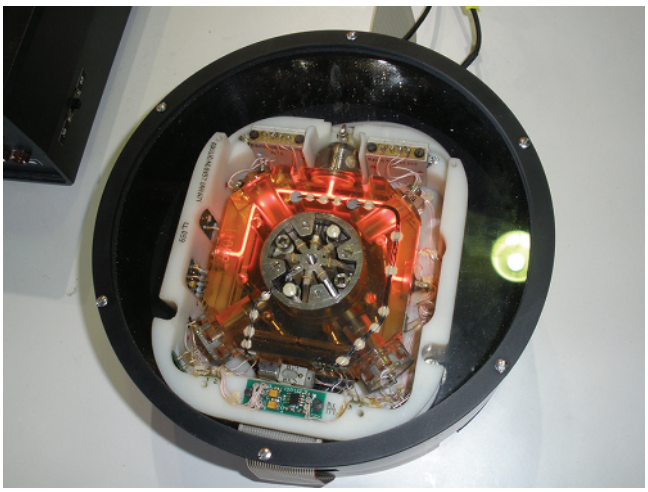


Figure 1.32 Single axis of a ring laser gyro. *Source:* Reproduced with permission of https://pl.wikipedia.org/wiki/%C5%BByroskoplaserowy#/media/File:RinglasergyroscopeatMAKS-2011airshow.jpg. Licensed under cc by sa 3.0 [5].

These are micromachined into Si and are exceedingly small and light with a three-axis accelerometer packaged onto a chip a few mm across. In principle, MEMS rate gyros could also be used on the INS platform but currently these have not achieved the required accuracy for INS. If this is achieved in the future, however, there is the prospect of an "INS on a chip" weighing a few grams that would fit into a matchbox. Current units have MEMS accelerometers and RLGs used as rate gyros on the INS platform.

One of the most significant changes that has occurred with INS in the last few years is a change in the philosophy of navigation from system-based (i.e. depending on a specific system like GNSS or INS) to performance-based. This involves combining all the available navigation sensors on an aircraft to achieve a specific navigation performance as discussed in the next section, thus current INS systems are combined with GNSS in a single navigation system. In this case the inertial measurement equipment is referred to as an Inertial Reference System (IRS) as its sensors have become part of a larger navigation system. In addition, the ADC can be combined with the IRS to produce an Air Data Inertial Reference System (ADIRS) in which the IRS has access to measurements of height, speed, Mach No, etc.

1.8.4 Combining Systems: Performance-Based Navigation (PBN) and Required Navigation Performance (RNP)

A modern airliner has a number of navigation sensors, including GPS, INS, and VOR, some of which may be triple redundant and all of which have different

levels of accuracy of position in different flight phases. For example, the position error with INS always increases with time (see Chapter 9), whereas for GPS it depends on the number and location of satellites in view and whether augmentation is available. For VOR tracking the position error decreases with decreasing distance to the VOR beacon. Since the early 1980s, Flight Management Computers (FMCs) have been a feature on the flight deck to manage navigation and integrate it with performance management and other functions [14]. Partly, FMCs were introduced to decrease crew workload as airlines started to move to two-crew operations. Thus, the output from all navigational sensors along with those of the ADC, engine performance indicators, etc. are available to the FMS. The navigation sensors are combined using a process known as Kalman filtering (see Section 9.9), which essentially means taking a weighted average of the positions provided by the different sensors. The weighting for a specific sensor increases with the accuracy of that sensor in the relevant flight phase, so, for example, INS would have a greater weighting at the beginning of a flight than at the end. The data can all be combined to give an accuracy parameter known as Actual Navigation Performance (ANP), defined as the radius of a circle centered on the computed current position, where the probability of the aeroplane remaining continuously inside the circle is 95% per flight hour.

This facilitates a new concept in Air Traffic Management (ATM) in which routes are not specified by traditional airways or approaches via beacons but by optimized curved paths in three dimensions with a specified Required Navigational Performance (RNP) for aircraft to follow that path. In its simplest form this defines the maximum allowed deviation from the path so the radius of the ANP circle must be less than the deviation. For example, RNP 4 specifies that the deviation is less than 4 nautical miles, which may be appropriate for enroute navigation, while RNP 0.1 would be used for an approach. The RNP path would be between two points that may be at different heights so that in addition to Lateral Navigation (LNAV), Vertical Navigation (VNAV) is also important. GNSS and INS both provide altitude information and this can be combined with the height reading from the ADC. A good proportion of the current fleet of commercial airliners has the avionics necessary for this change and the ANP along with the RNP for a particular phase of flight is displayed on the PFD as shown in Figure 1.33a. The bars at the bottom of the display show the difference between the ANP and the RNP with reference to Figure 1.33b and the deviation from the ideal LNAV path is indicated by the small triangular pointer. The gap between the bars provides some indication of the leeway available for deviation to the flight crew, for example, if they needed to fly around a thunderstorm, deviations up to the edge of the bar are acceptable without calling ATC for a clearance. ANP–RNP bars are also displayed at the side of the PFD for VNAV.

Figure 1.34a and b illustrate the difference between a conventional approach using radio beacons and a curved three-dimensional approach RNP. In the latter case, there is a saving in distance traveled and fuel used while maintaining safe

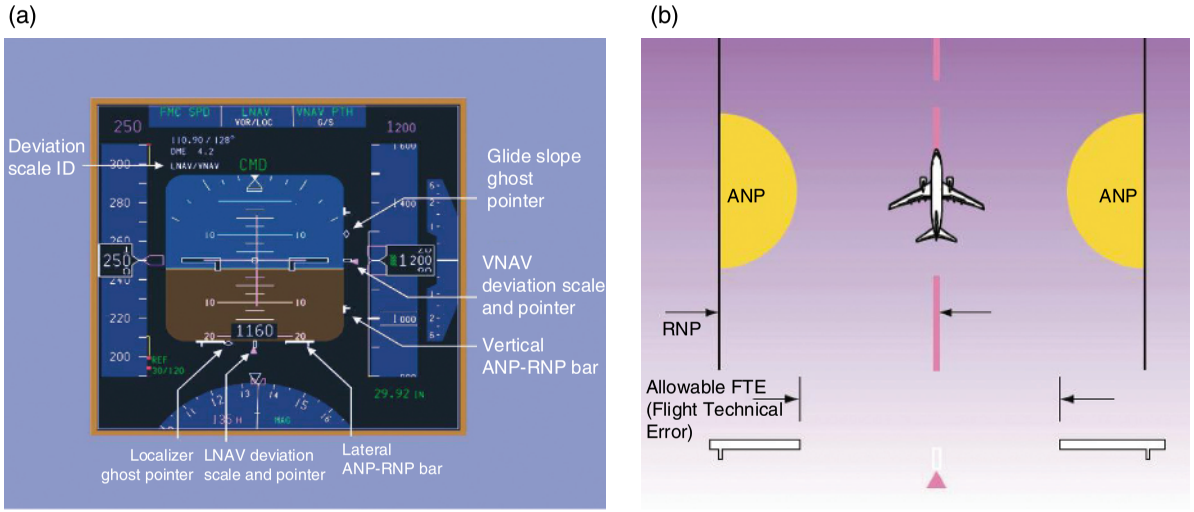


Figure 1.33 (a) PFD showing ANP–RNP for both LNAV (bottom of the display) and VNAV (right-hand side of the display. *Source:* Reproduced with permission of Aero Magazine, 2001. (b) Illustration of the information provided by the ANP–RNB bars on the PFD. The deviations from the ideal lateral and vertical paths are also shown providing some indication to the crew about how much deviation is permissible for the current RNP. *Source:* Reproduced with permission of Aero magazine 2001 [15].

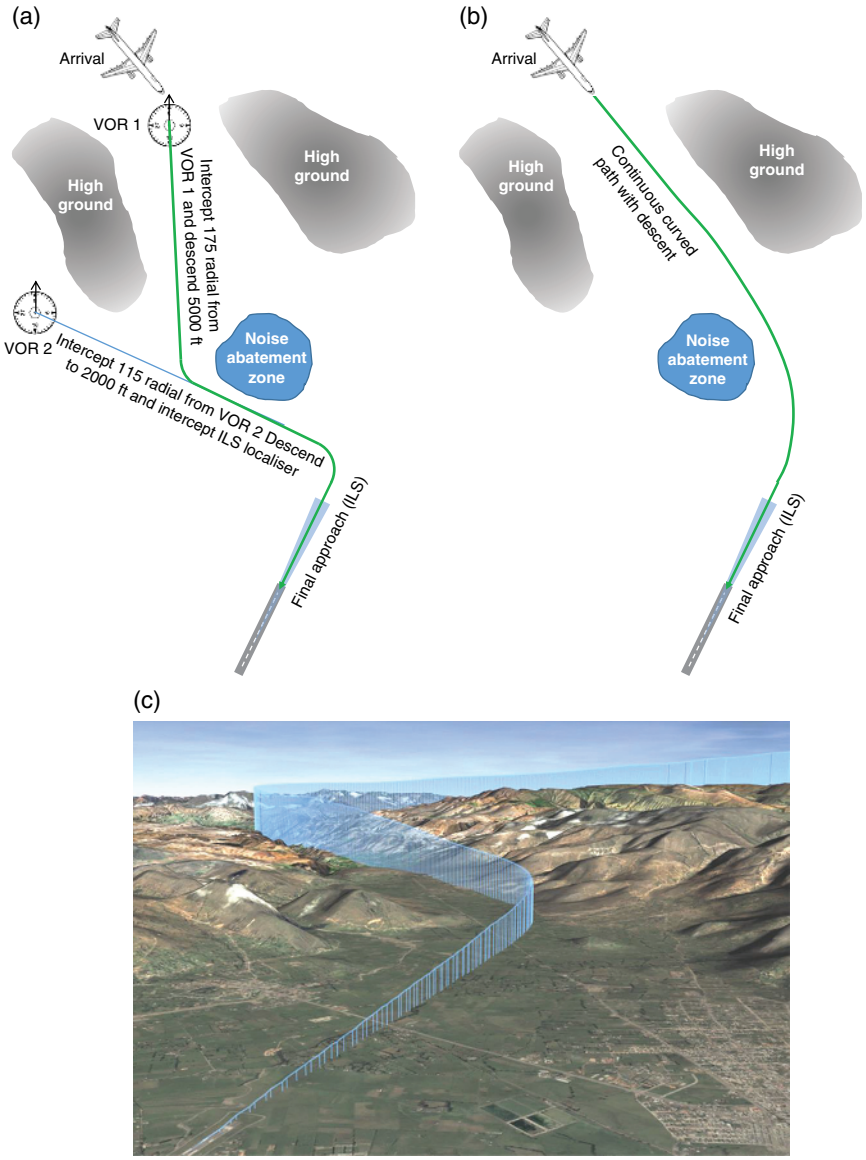


Figure 1.34 (a) Example of a conventional approach procedure using specific radials from VORs to line up with a runway. (b) Curved three-dimensional RNP approach to line up on the same runway. (c) RNP approach to Cajamarca airport, Peru. *Source:* Reproduced with permission of https://en.wikipedia.org/wiki/Requirednavigationperformance#/media/File:RNPTrack3D.png. Licensed under CC BY-SA 3.0 [5].

separation from terrain and avoiding overflying noise abatement zones. RNP approaches are particularly suited to mountainous areas as curved tracks can be defined to safely navigate obstacles. The first RNP approach tested was in 1996 by Alaskan Airlines into Juneau airport, which has a challenging local terrain and an example of the track for an RNP approach to Cajamarca airport, Peru, is shown in Figure 1.34c. Such tracks will be stored on the navigation database and can be made active for the autopilot to follow.

1.9 Development of Auto Flight Control Systems

Like many of the topics in this chapter, the history of the autopilot begins in the nineteenth century before the first powered flight. As pointed out by McRuer and Graham [16], an important mind shift occurred with the Wright brothers aeroplane, which had well thought out controls for correcting changes in pitch, roll, and yaw. They effectively abandoned the search for a design with inherent stability in all axes and used feedback from the pilot to induce stability in the aeroplane. The pilot uses sensors to detect changes, calculates the response required to bring the aircraft back to level flight, applies the necessary force, which is amplified to move the control surfaces to produce the response. All these functions can be reproduced by a machine and the important point is that since the dawn of aviation someone or something needs to be in control of the aircraft. Since before the Wright brothers, inventors had experimented with the something, that is, a machine that could maintain an aircraft in stable flight. All early autopilots (although the term was not used till later) were based on the rigidity of the plane of rotation of gyroscopes, which, as pointed out in Section 1.4, must have been noted before recorded history.

The basic mechanism by which a gyroscope can be used to stabilize the flight attitude of an aircraft is illustrated in Figure 1.35 for just one axis (roll). The gyroscope cage is in a single gimbal, which is attached to the aircraft frame and the rotor is set spinning with its plane parallel to the wings (and to the horizon). Figure 1.35b shows what happens when the aircraft rolls by using a pointer attached to the gimbal and a scale attached to the rotor cage. The gyroscope maintains its original plane of rotation parallel to the horizon and the pointer indicates the roll angle, which is in fact one of the motions of an artificial horizon (see Section 3.1.9). Figure 1.35c shows the rolled attitude as seen from an observer in the aircraft, that is, the gimbal appears to have stayed stationary while the rotor has tilted. This movement in the rotor can be used to actuate a device that controls the ailerons and rotates the wings back to parallel to the horizon. Having two similar gyroscopes set up to rotate in orthogonal directions can also control the pitch and yaw to maintain straight and level flight.

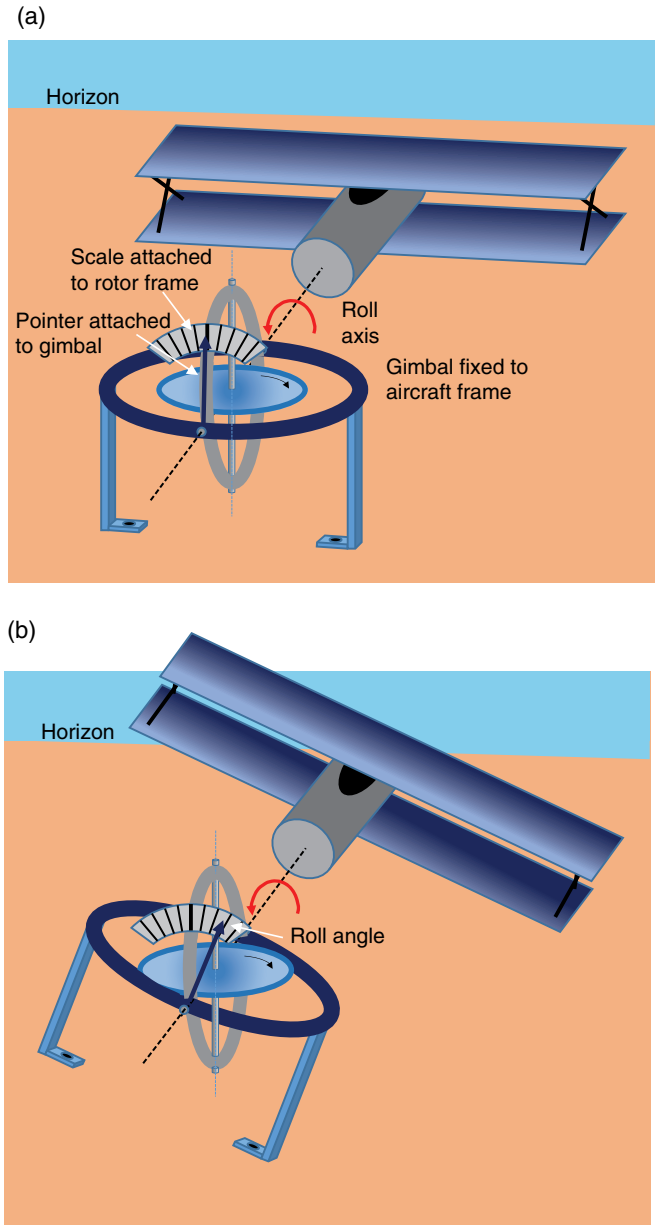


Figure 1.35 Gyroscope set up to control roll. (a) The caged rotor is in a single gimbal attached to the aircraft frame with a pointer attached to the gimbal and a scale attached to the cage. The gyro is set spinning with its plane of rotation parallel to the wings and the horizon (b) When the aircraft rolls the rotor plane stays parallel to the horizon and the pointer shows the roll angle (one axis of an artificial horizon). (c) Viewed from within the aircraft the plane of the rotor rotates as the aeroplane rolls and this movement can control an actuator that rolls the wings back to be parallel to the horizon. The pitch and yaw axes can similarly be controlled by orthogonal gyroscopes.

(c)

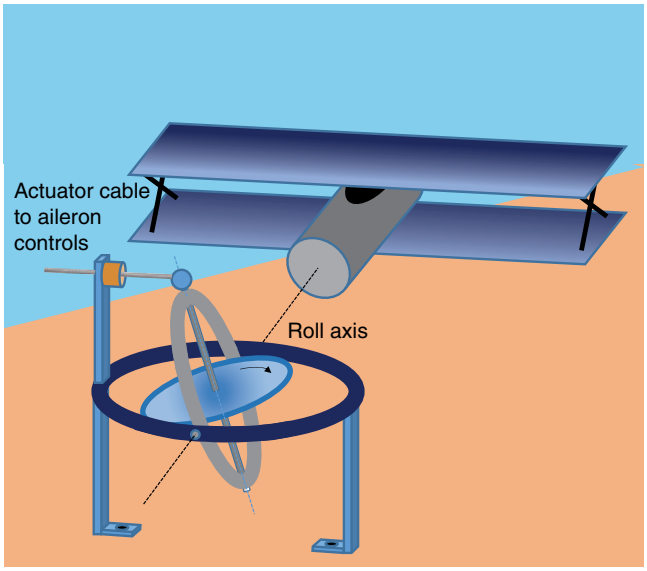


Figure 1.35 (Continued)

The first attempt to use this as method of control was by the US-born inventor Hiram Maxim in 1891, who used a gyroscope with a steam driven rotor to produce pitch stability in his heavier than air flying machine. The movement of the rotor relative to the aircraft longitudinal axis was used to operate a valve that allowed steam into a piston that controlled the elevators. The invention did not develop much beyond the drawing board due to the aeroplane being destroyed during early tests after which Maxim appeared to lose interest and moved on to other things. It was left to Lawrence Sperry to bring the gyroscopic control of flight to the attention of the world in 1914.

Lawrence, the son of Elmer Sperry, who founded the Sperry gyroscope company in 1910, was a talented inventor and by the age of 17 had built his first aeroplane. He was also a natural pilot and taught himself to fly but decided to enroll in the flying school run by Glenn Curtiss to obtain his Federal aeronautics license in 1913. This was the 11th to be issued and he was the youngest holder of a license in the United States. He had the free run of the workshop at the Curtiss school and immediately began work on a method to stabilize aircraft in flight based on gyroscopes. To the young Sperry the rigidity of gyroscopes to maintain a reference plane seemed a natural characteristic to exploit to maintain an aircraft at a constant flight attitude. He used an independent gyroscope to control each of the aircraft rotations, that is, roll, pitch, and yaw.

Although most aircraft flying in 1913 had adopted the control surfaces we are familiar with today, that is, ailerons, elevators, and rudders, the pilot controls depended on the manufacturer. Fortunately for Sperry it was about this time that the aviation industry standardized on the most ergonomic system to emerge, which was that designed by the French plane maker SPAD owned by Armand Deperdussin. This is the system that is still used today in which a stick or control yoke is moved backwards and forward to control the elevators (pitch), side to side (or wheel left and right) to control the ailerons (roll), and foot pedals for the rudder (yaw). Until recently the tendency has been for large aircraft and cabin GA aircraft to have control yokes (wheels) while sticks were used on open cockpits, fighters, and aerobatics planes. This has changed recently with Airbus airliners and Cirrus and Diamond GA aircraft adopting sticks for manual control.

With a universal control system in place it became a lot simpler for Sperry to design his autopilot and he finally managed to package it in a box the size of a cabin bag weighing 18 kg. After testing and refining in the United States it was first displayed to the world in a spectacular fashion at the Concours de la Securité en Aéroplane (Aeroplane Safety Competition) held in Paris on 18 June 1914. Not content with just showing a hands-off flyby, on the second pass, Sperry's French mechanic climbed out onto the wing while Sperry held his hands in the air (Figure 1.36a). On the final pass by the judges stand, both men climbed out onto the wings and waved to the ecstatic crowds as the plane flew on straight and level by itself. The slight rolls that developed as they climbed out were quickly corrected as the plane rolled back to wings level. It was a hard act to follow and Sperry won the prize of 50 000 francs and at the same time became a renowned inventor. The Sperry autopilot was used in the first cruise missile, the Kettering Bug built in 1918 (Figure 1.36b), of which

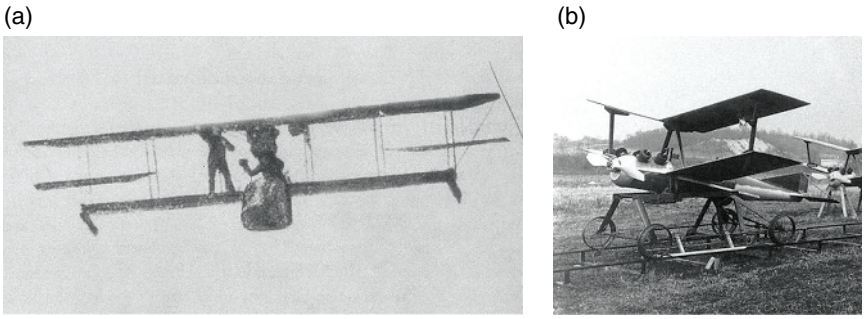


Figure 1.36 (a) Lawrence Sperry and his mechanic demonstrating the gyroscopic autopilot in Paris in 1914. (b) The Kettering bug cruise missile in 1918, which was fitted with the Sperry autopilot. *Source:* (a) HistoricWings.com and (b) HistoricWings.com.

many were built but never used in action. A gyroscopic stabilization mechanism was, however, used on the German V1 cruise missile in World War 2.

From 1914 to post World War 2, gyroscopes remained the basic sensing mechanism for control but they became more compact and the actuator systems that moved the control surfaces became increasingly sophisticated and included pneumatic servos. In addition, the autopilots started to take additional inputs from the compass and the pressure altimeter so that they could maintain an aircraft on a specific course and height. In 1930, a US Army Air Corps plane was kept at a constant altitude and heading for three hours [17]. The Royal Aircraft Establishment (RAE) in the United Kingdom were also very active in autopilot development, though much of it was kept secret and in 1930 they demonstrated automatic altitude and heading control of an aircraft for over 400 miles [18]. These developments were taking place when the length of commercial flights was increasing and it was recognized that pilots required assistance for flights that were several hours long, especially in nonvisual conditions.

In 1932, the prototype of what was to become the standard autopilot for commercial airliners, the Sperry A2, was developed. By then, airliners were already fitted with an extensive set of gyroscopic instruments (see Section 1.4) to cope with flight in low visibility conditions and one of the innovative steps in the A2 design was to use the instruments to supply the gyroscopic sensing. This saved weight as there was no need to have an independent set of gyros and the autopilots started to resemble modern systems in which sense inputs come from the FMS. The timing of the A2 coincided perfectly with the needs of commercial aviation and its profile was greatly enhanced by Wiley Post in his record-breaking round the world flight in a Lockheed Vega aircraft, the "Winnie Mae" in 1933 (Figure 1.37a). Post had seen the A2 prototype at the Sperry factory and insisted

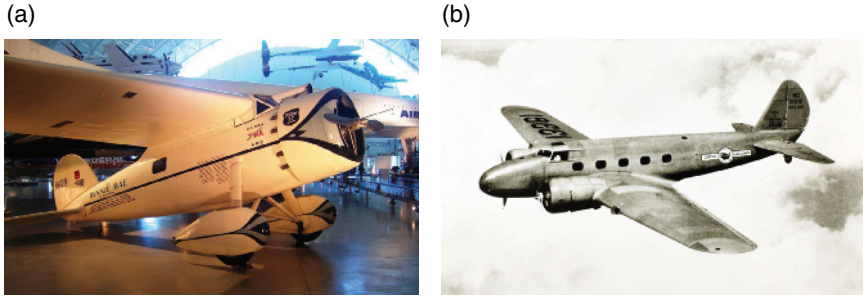


Figure 1.37 (a) Lockheed Vega "Winnie Mae" piloted by Wiley post on his record-breaking round the world flight in 1933 used the Sperry A2 autopilot to rest. *Source:* (a) https://en.wikipedia.org/wiki/WileyPost#/media/File:NASM-LockheedVega-WinnieMae.jpg. Licensed under CC BY-SA 3.0. (b) Boeing 247 airliner was the first to be fitted with an A2 autopilot as standard.

on having one fitted to his plane, which he used extensively to get some rest on his 7-day 19-hour flight. The A2 was first introduced into commercial service on the Boeing 247 (Figure 1.37b) in 1934 and was quickly in widespread use by the airlines. It used pneumatics to sense and amplify the indications of the sensors (gyroscopes) and hydraulics for the actuators on the control surfaces; however, the trend was to move to all-electric systems, that is, electrically driven gyroscopes and sensors, power amplification, and electric motors to drive the control surfaces. Other companies including Honeywell and Bendix in the United States and Siemens, Askania, and Anschütz in Germany entered the field. Similar rapid developments were also being made by the RAE in the United Kingdom though this continued to remain mostly secret and very little was published.

During World War 2, more development steps in autopilots emerged in Germany, including yaw damping to remove some fundamental aerodynamic instabilities of aircraft and the principle of rate-rate control. This is where rate gyros are used to measure the rate of rotation in roll, pitch, or yaw and the correcting control surface is moved at a rate that is proportional providing more positive control. In the United States the move to all electric autopilots was realized and also the emergence of a new level of control, that is, the ability of autopilots to maneuver aircraft rather than maintain a set attitude.

All the mechanisms presented so far would be described in modern terminology as *inner loop systems*, that is, they contain a single feedback loop that *maintains* a specific flight parameter (height, heading, etc.). It was becoming clear that it would be necessary to introduce a higher level of control, which on reception of some additional input could *change* one of the flight parameters and then let the inner loop system maintain it at the new value. An example is obtaining information from a bombsight, change to a new heading, and maintain it till new information is received. Such a control method is referred to as an *outer loop system* and is the basis of how the autopilot in a modern airliner can be made to follow a flight plan. Another example of an outer loop system is commands from the pilot controls for the autopilot to change to a new flight attitude, which forms the basis of fly-by-wire systems,

The sophistication that had been reached by autopilots was demonstrated convincingly in 1947 when a US Air Force C47 Skymaster flew from Stephenville, Newfoundland to Brize Norton in England without any pilot input from brake release at takeoff to brakes on after landing. This impressive feat was achieved with the latest technology of the time including a Sperry A-12 autopilot with approach coupler and Bendix automatic throttle control. The commands to the autopilot were input from punched cards interpreted by an IBM electronic controller so as a system it had the same capabilities as a modern autopilot commanded by a FMC. Considering that landing an aircraft is the most difficult task a human pilot performs, this flight seemed to indicate that

there was nothing an autopilot could not achieve and indeed the majority of commercial flights today are carried out mainly by the autopilot.

Emerging from the war, the autopilot systems had clearly become more sophisticated but some deficiencies still remained including their tendency to over-control and induce oscillations. As pointed out by Mcruer, Graham, and Ashkensas [19], the theoretical basis on how to remove this kind of deficiency had been available for some time but the merging of theory and the practical application of autopilots did not occur till after World War 2. For example, the Proportional-Integral-Derivative (PID) control method used in a range of industrial control processes to achieve the optimum rate of convergence of a control parameter with a desired value followed by optimum stability was published in 1922 [20]. This merging of theory and practical application in the late 1940s and early 1950s coincided with the development of computers that could be used to solve the complex equations involved so the development of autopilots accelerated.

The next major development was routine landing by the automatic pilot (autoland) in commercial flights. This involved not only the development of equipment but rules and procedures such as safety minima. Although the principle of autoland had been demonstrated in 1947 by the US Air Force, this was with nonstandard equipment not generally available on airliners. In 1945, the British Government set up the Blind Landing Experimental Unit (BLEU) to carry out research and development on low visibility landings. This was an imperative for the nationalized airlines, BOAC set up in 1940 and BEA in 1946 since the large amount of air pollution created by coal fires coupled with damp weather could close London airports for days on end. Autoland was developed by BLEU initially for RAF aircraft and in the early 1960s for BEA's Trident fleet. It used a triple-redundant control with three independent processing channels and if one failed the other two would "out-vote" it and provide the output for the controls. As already described in Section 1.7.3, the technology came to maturity with the first fully automatic landing on a commercial flight with passengers at Heathrow airport in a British European Airways Trident in 1965.

Thus, the fundamental methods and hardware for autopilots to control the aircraft through every phase of flight were in place by the 1960s. From then on, the increasing sophistication of the FMS in the cockpit driven by ever-more powerful digital computers meant that new modes could be invoked, for example, terrain following and terrain avoidance. In 1972, the first digital fly by wire (FBW) system without a mechanical backup was tested by NASA on a Crusader F-8C. In this type of servo-mechanism there is no direct mechanical or hydraulic coupling between the flight controls and the control surfaces but the flight controls send electronic signals to the same servos that the autopilot uses to move the control surfaces. The 1970s also saw the development of full authority digital engine control (FADEC) where engine control is delegated

Table 1.1 Typical modes of the autopilot in a commercial aircraft.

No.	Mode	Action
1	Heading	Follows a selected heading (e.g. 280°).
2	LNAV	Follows the lateral route entered in the FMS.
3	VOR/LOC	Follows a selected track to or from a VOR or ILS localizer entered manually or selected by the FMS.
4	Altitude hold	Holds the altitude while pilot maintains lateral control.
5	Vertical speed	Maintains a specified vertical speed (climb or descent) till a selected altitude is intercepted.
6	Level change	Climbs or descends to the selected altitude while maintaining the selected speed.
7	VNAV	Follows the vertical component of the route entered into the FMS.
8	ILS/approach	Follows the glide slope and localizer to the runway and carries out autoland if the necessary systems are available.

to an electronic subsystem that takes care of all the details in demanding a given power from the engine. It has inputs from the ADC and FMS so that when a given power is requested, the system controls all the engine settings required, for example, fuel flow, intake configuration, propeller pitch, etc. to deliver that power. This makes engine control much simpler for both human pilots and autopilots as they just need to specify a single number – the percentage power required and the FADEC does the rest. It is an example of the distributed processing that is found on a modern aircraft.

The autopilot modes selectable by the pilot in a typical modern airliner are shown in Table 1.1 and it is evident that the pilot can demand all the necessary actions during the course of a normal flight to be carried out by the autopilot. If the autopilot is switched to LNAV and VNAV, then it will follow the flight pattern entered into the FMS (usually by airline operations) but the extra flexibility of the modes shown in Table 1.1 allow the crew to make changes easily in response to requests from ATC. The FMS menu system is also carefully designed so that modifications can be made to the flight plan with minimum workload.

The increase in sophistication of autopilots since the demonstration by Lawrence Sperry in Paris a century ago is truly impressive. Most commercial flights are conducted mainly by the autopilot and it is known that this saves fuel and increases passenger comfort. It must be borne in mind, however, that even today there are situations, either due to turbulence or malfunctioning sensors that an autopilot cannot deal with and it drops out leaving it to the human pilots to complete the flight.

References

1 Roger Ford (2000). The risks of travel. *Modern Railways* (October) (article cites figures based on UK Department of the Environment, Transport and the Regions (DETR) survey).
2 http://ethw.org/Milestones:First_Blind_Takeoff,_Flight_and_Landing,_1929 (accessed 11 June 2018).
3 Benniwitz, K. (1922). *Flugzeuginstrumente*. Berlin: Richard Carl Schmidt and Co.
4 Wing Cdr. Roderic Hill (2005). *The Baghdad Air Mail*. Nonsuch Publishing Ltd.
5 Creative commons (1953). http://creativecommons.org/licenses/by-sa/3.0/legalcode (accessed 11 June 2018).
6 www.airbattle.co.uk/b_research_3.html (accessed 11 June 2018).
7 http://www.aps.org/publications/apsnews/200604/history.cfm (accessed 11 June 2018).
8 http://creativecommons.org/licenses/by/2.0/legalcode (accessed 11 June 2018).
9 Johnson, R. (2003). Blind flying on the beam: aeronautical communications, navigation and surveillance: its origins and the politics of technology, part III: emerging technologies, the radio range, the radio beacon and the visual indicator. *Journal of Air Transportation* **8**: 79–104.
10 https://timeandnavigation.si.edu/multimedia-asset/gee-chart-reims-chain-december-1944-11000000-scale (accessed 11 June 2018).
11 Helfrick, A. (2015). The centennial of avionics: our 100 year trek to performance-based navigation. *IEEE A & E Systems Magazine, Sept* **30** (9): 36–45.
12 Collinson, R.P.G. (2003). *Introduction to Avionics Systems*, 226. Springer.
13 Savage, P.G. (2013). Blazing gyros – the evolution of strapdown inertial navigation technology for aircraft. *AIAA Journal of Guidance, Control and Dynamics* **36**: 637–655.
14 Miller, S. (2009). Contribution of flight systems to performance-based navigation. *Boeing Aero Magazine* QTR_02.09, pp. 20–28.
15 Carriker, M., Hilby, D., Houck, D., and Rolan Shomber, H. (2001). Lateral and vertical navigation deviation displays. *Boeing Aero Magazine*, No. 16 (October), pp. 29–35.
16 McRuer, D. and Graham, D. (1981). Eighty years of flight control: triumphs and pitfalls of the systems approach. *Journal of Guidance and Control* **4**: 353.
17 Now – the automatic pilot (1930). *Popular Science Monthly* (February), p. 22.
18 Robot air pilot keeps plane on true course (1930). *Popular Mechanics* (December), p. 950.
19 Mcruer, D.T., Graham, D., and Ashkensas, I. (1973). *Aircraft Dynamics and Automatic Control*, 6. Princeton University Press.
20 Minosrsky, N. (1922). Directional stability of automatically steered bodies. *Journal of American Society of Naval Engineers* **34**: 280–309.