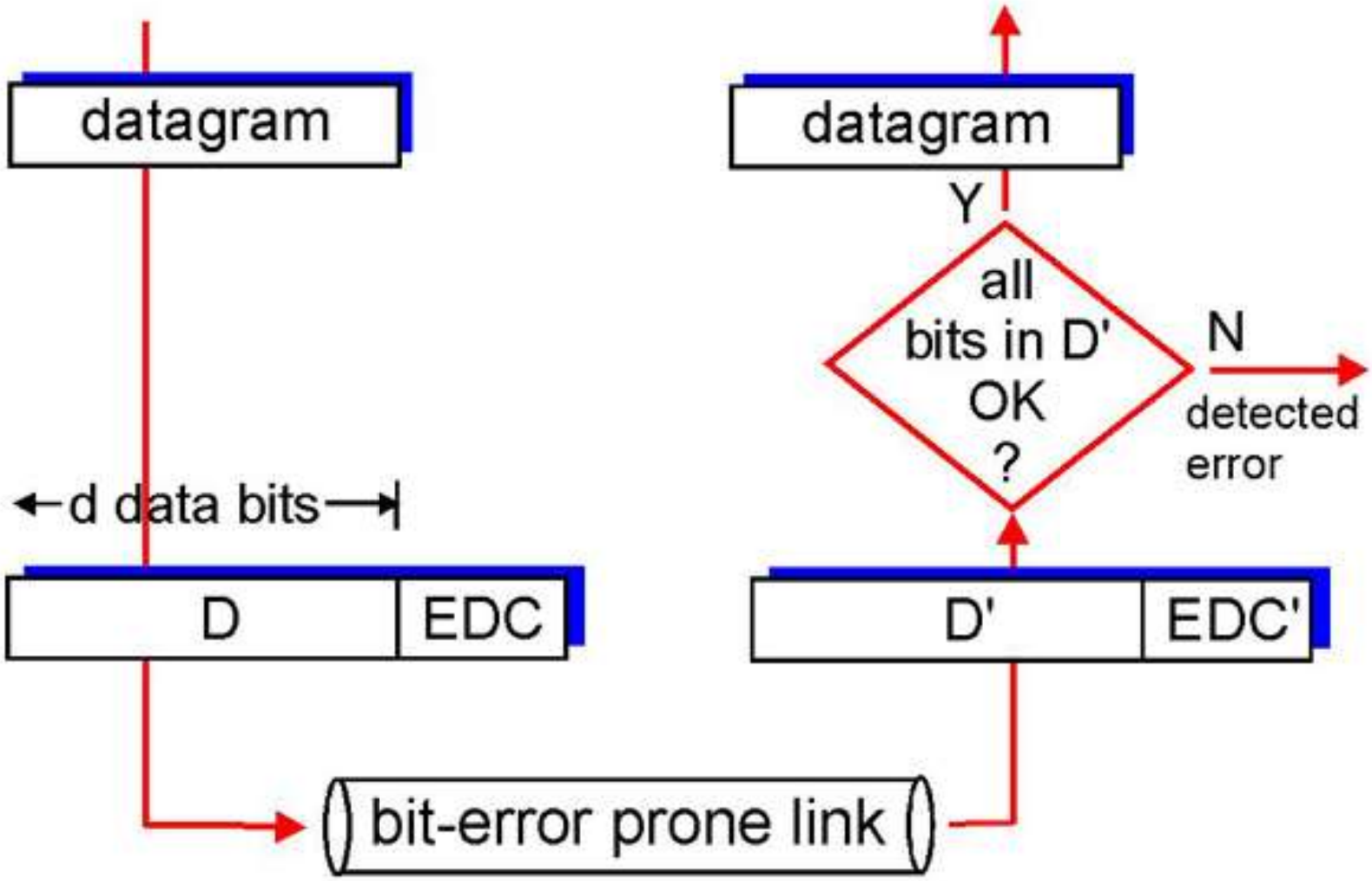
Fault Tolerant Design, Error Detection and Correction



Stephaine Hatcher

Porsha Whitman

First Edition, 2012

ISBN 978-81-323-0882-9

Published by:
Academic Studio
4735/22 Prakashdeep Bldg,
Ansari Road, Darya Ganj,
Delhi - 110002
Email: info@wtbooks.com

Table of Contents

Chapter-1

Fault-Tolerant Design

In engineering, **fault-tolerant design**, also known as **fail-safe design**, is a design that enables a system to continue operation, possibly at a reduced level (also known as graceful degradation), rather than failing completely, when some part of the system fails. The term is most commonly used to describe computer-based systems designed to continue more or less fully operational with, perhaps, a reduction in throughput or an increase in response time in the event of some partial failure. That is, the system as a whole is not stopped due to problems either in the hardware or the software. An example in another field is a motor vehicle designed so it will continue to be drivable if one of the tires is punctured. A structure is able to retain its integrity in the presence of damage due to causes such as fatigue, corrosion, manufacturing flaws, or impact.

Components

If each component, in turn, can continue to function when one of its subcomponents fails, this will allow the total system to continue to operate, as well. Using a passenger vehicle as an example, a car can have "run-flat" tires, which each contain a solid rubber core, allowing them to be used even if a tire is punctured. The punctured "run-flat" tire may be used for a limited time at a reduced speed.

Redundancy

This means having backup components which automatically "kick in" should one component fail. For example, large cargo trucks can lose a tire without any major consequences. They have many tires, and no one tire is critical (with the exception of the front tires, which are used to steer).

When to use

Providing fault-tolerant design for every component is normally not an option. In such cases the following criteria may be used to determine which components should be fault-tolerant:

- **How critical is the component?** In a car, the radio is not critical, so this component has less need for fault-tolerance.

- **How likely is the component to fail?** Some components, like the drive shaft in a car, are not likely to fail, so no fault-tolerance is needed.

- **How expensive is it to make the component fault-tolerant?** Requiring a redundant car engine, for example, would likely be too expensive both economically and in terms of weight and space, to be considered.

An example of a component that passes all the tests is a car's occupant restraint system. While we do not normally think of the *primary* occupant restraint system, it is gravity. If the vehicle rolls over or undergoes severe g-forces, then this primary method of occupant restraint may fail. Restraining the occupants during such an accident is absolutely critical to safety, so we pass the first test. Accidents causing occupant ejection were quite common before seat belts, so we pass the second test. The cost of a redundant restraint method like seat belts is quite low, both economically and in terms or weight and space, so we pass the third test. Therefore, adding seat belts to all vehicles is an excellent idea. Other "supplemental restraint systems", such as airbags, are more expensive and so pass that test by a smaller margin.

Examples

Hardware fault-tolerance sometimes requires that broken parts can be swapped out with new ones while the system is still operational (in computing known as *hot swapping*). Such a system implemented with a single backup is known as **single point tolerant**, and represents the vast majority of fault-tolerant systems. In such systems the mean time between failures should be long enough for the operators to have time to fix the broken devices (mean time to repair) before the backup also fails. It helps if the time between failures is as long as possible, but this is not specifically required in a fault-tolerant system.

Fault-tolerance is notably successful in computer applications. Tandem Computers built their entire business on such machines, which used single point tolerance to create their **NonStop** systems with uptimes measured in years.

Fail-safe architectures may encompass also the computer software, for example by process replication (computer science).

Disadvantages

Fault-tolerant design's advantages are obvious, while many of its disadvantages are not:

- **Interference with fault detection in the same component.** To continue the above passenger vehicle example, it may not be obvious to the driver when a tire has been punctured, with either of the fault-tolerant systems. This is usually

handled with a separate "automated fault detection system". In the case of the tire, an air pressure monitor detects the loss of pressure and notifies the driver. The alternative is a "manual fault detection system", such as manually inspecting all tires at each stop.

- **Interference with fault detection in another component.** Another variation of this problem is when fault-tolerance in one component prevents fault detection in a different component. For example, if component B performs some operation based on the output from component A, then fault-tolerance in B can hide a problem with A. If component B is later changed (to a less fault-tolerant design) the system may fail suddenly, making it appear that the new component B is the problem. Only after the system has been carefully scrutinized will it become clear that the root problem is actually with component A.

- **Reduction of priority of fault correction.** Even if the operator is aware of the fault, having a fault-tolerant system is likely to reduce the importance of repairing the fault. If the faults are not corrected, this will eventually lead to system failure, when the fault-tolerant component fails completely or when all redundant components have also failed.

- **Test difficulty.** For certain critical fault-tolerant systems, such as a nuclear reactor, there is no easy way to verify that the backup components are functional. The most infamous example of this is Chernobyl, where operators tested the emergency backup cooling by disabling primary and secondary cooling. The backup failed, resulting in a core meltdown and massive release of radiation.

- **Cost.** Both fault-tolerant components and redundant components tend to increase cost. This can be a purely economic cost or can include other measures, such as weight. Manned spaceships, for example, have so many redundant and fault-tolerant components that their weight is increased dramatically over unmanned systems, which don't require the same level of safety.

- **Inferior components.** A fault-tolerant design may allow for the use of inferior components, which would have otherwise made the system inoperable. While this practice has the potential to mitigate the cost increase, use of multiple inferior components may lower the reliability of the system to a level equal to, or even worse than, a comparable non-fault-tolerant system.

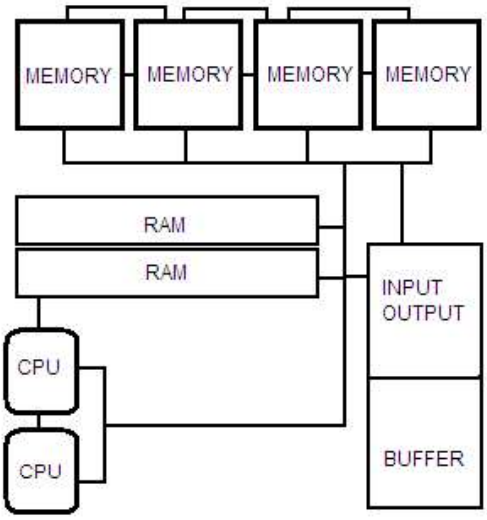
Related terms

There is a difference between fault-tolerance and systems that rarely have problems. For instance, the Western Electric crossbar systems had failure rates of two hours per forty years, and therefore were highly *fault resistant*. But when a fault did occur they still stopped operating completely, and therefore were not *fault-tolerant*.

Chapter-2

Fault-Tolerant Computer System and Byzantine Fault Tolerance

Fault-tolerant computer system



A conceptual design of a segregated-component fault-tolerant computer design

Fault-tolerant computer systems are systems designed around the concepts of fault tolerance. In essence, they have to be able to keep working to a level of satisfaction in the presence of faults.

Types of fault tolerance

Most fault-tolerant computer systems are designed to be able to handle several possible failures, including hardware-related faults such as hard disk failures, input or output device failures, or other temporary or permanent failures; software bugs and errors; interface errors between the hardware and software, including driver failures; operator errors, such as erroneous keystrokes, bad command sequences, or installing unexpected software; and physical damage or other flaws introduced to the system from an outside source .

Hardware fault-tolerance is the most common application of these systems, designed to prevent failures due to hardware components. Typically, components have multiple backups and are separated into smaller "segments" that act to contain a fault, and extra redundancy is built into all physical connectors, power supplies, fans, etc. . There are special software and instrumentation packages designed to detect failures, such as fault masking, which is a way to ignore faults by seamlessly preparing a backup component to execute something as soon as the instruction is sent, using a sort of voting protocol where if the main and backups don't give the same results, the flawed output is ignored.

Software fault-tolerance is based more around nullifying programming errors using real-time redundancy, or static "emergency" subprograms to fill in for programs that crash. There are many ways to conduct such fault-regulation, depending on the application and the available hardware..

History

The first known fault-tolerant computer was SAPO, built in 1951 in Czechoslovakia by Antonin Svoboda. Its basic design was magnetic drums connected via relays, with a voting method of memory error detection. Several other machines were developed along this line, mostly for military use. Eventually, they separated into three distinct categories: machines that would last a long time without any maintenance, such as the ones used on NASA space probes and satellites; computers that were very dependable but required constant monitoring, such as those used to monitor and control nuclear power plants or supercollider experiments; and finally, computers with a high amount of runtime which would be under heavy use, such as many of the supercomputers used by insurance companies for their probability monitoring.

Most of the development in the so called LLNM (Long Life, No Maintenance) computing was done by NASA during the 1960s, in preparation for Project Apollo and other research aspects. NASA's first machine went into a space observatory, and their second attempt, the JSTAR computer, was used in Voyager. This computer had a backup of memory arrays to use memory recovery methods and thus it was called the JPL Self-Testing-And-Repairing computer. It could detect its own errors and fix them or bring up redundant modules as needed. The computer is still working today.

Hyper-dependable computers were pioneered mostly by aircraft manufacturers, nuclear power companies, and the railroad industry in the USA. These needed computers with massive amounts of uptime that would fail gracefully enough with a fault to allow continued operation, while relying on the fact that the computer output would be constantly monitored by humans to detect faults. Again, IBM developed the first computer of this kind for NASA for guidance of Saturn V rockets, but later on BNSF, Unisys, and General Electric built their own.

In general, the early efforts at fault-tolerant designs were focused mainly on internal diagnosis, where a fault would indicate something was failing and a worker could replace it. SAPO, for instance, had a method by which faulty memory drums would emit a noise

before failure . Later efforts showed that, to be fully effective, the system had to be self-repairing and diagnosing – isolating a fault and then implementing a redundant backup while alerting a need for repair. This is known as N-model redundancy, where faults cause automatic fail safes and a warning to the operator, and it is still the most common form of level one fault-tolerant design in use today.

Voting was another initial method, as discussed above, with multiple redundant backups operating constantly and checking each other's results, with the outcome that if, for example, four components reported an answer of 5 and one component reported an answer of 6, the other four would "vote" that the fifth component was faulty and have it taken out of service. This is called M out of N majority voting.

Historically, motion has always been to move further from N-model and more to M out of N due to the fact that the complexity of systems and the difficulty of ensuring the transitive state from fault-negative to fault-positive did not disrupt operations.

Fault tolerance verification and validation

The most important requirement of design in a fault tolerant computer system is making sure it actually meets its requirements for reliability. This is done by using various failure models to simulate various failures, and analyzing how well the system reacts. These statistical models are very complex, involving probability curves and specific fault rates, latency curves, error rates, and the like. The most commonly used models are HARP, SAVE, and SHARPE in the USA, and SURF or LASS in Europe.

Fault tolerance research

Research into the kinds of tolerances needed for critical systems involves a large amount of interdisciplinary work. The more complex the system, the more carefully all possible interactions have to be considered and prepared for. Considering the importance of high-value systems in transport, utilities and the military, the field of topics that touch on research is very wide: it can include such obvious subjects as software modeling and reliability, or hardware design, to arcane elements such as stochastic models, graph theory, formal or exclusionary logic, parallel processing, remote data transmission, and more.

Byzantine fault tolerance

Byzantine fault tolerance is a sub-field of fault tolerance research inspired by the **Byzantine Generals' Problem**, which is a generalized version of the Two Generals' Problem.

The object of Byzantine fault tolerance is to be able to defend against *Byzantine failures*, in which components of a system fail in arbitrary ways (i.e., not just by stopping or crashing but by processing requests incorrectly, corrupting their local state, and/or producing incorrect or inconsistent outputs.). Correctly functioning components of a Byzantine fault tolerant system will be able to correctly provide the system's service assuming there are not too many Byzantine faulty components.

Byzantine failures

A **Byzantine fault** is an arbitrary fault that occurs during the execution of an algorithm by a distributed system. It encompasses both **omission failures** (e.g., crash failures, failing to receive a request, or failing to send a response) and **commission failures** (e.g., processing a request incorrectly, corrupting local state, and/or sending an incorrect or inconsistent response to a request.) When a Byzantine failure has occurred, the system may respond in any unpredictable way, unless it is designed to have Byzantine fault tolerance.

For example, if the output of one function is the input of another, then small round-off errors in the first function can produce much larger errors in the second. If the second function were fed into a third, the problem could grow even larger, until the values produced are worthless. Another example is in compiling source code. One minor syntactical error early on in the code can produce large numbers of perceived errors later, as the parser of the compiler gets out-of-phase with the lexical and syntactic information in the source program. Such failures have brought down major Internet services. For example, in 2008 Amazon S3 was brought down for several hours when a single-bit hardware error propagated through the system, and in 2009 the Ma.gnolia bookmark sharing website was shuttered after a file system error gradually corrupted the system's database beyond recovery.

In a Byzantine fault tolerant (BFT) algorithm, steps are taken by processes, the logical abstractions that represent the execution path of the algorithms. A faulty process is one that at some point exhibits any of the above failures. A process that is not faulty is correct.

The Byzantine failure assumption models real-world environments in which computers and networks may behave in unexpected ways due to hardware failures, network congestion and disconnection, as well as malicious attacks. Byzantine failure-tolerant algorithms must cope with such failures and still satisfy the specifications of the

problems they are designed to solve. Such algorithms are commonly characterized by their resilience t, the number of faulty processes with that an algorithm can cope.

Many classic agreement problems, such as the Byzantine Generals' Problem, have no solution unless $n > 3t$, where n is the number of processes in the system. In other words, the algorithm can ensure correct operation only, if fewer than one third of the processes are faulty.

Origin

Byzantine refers to the Byzantine Generals' Problem, an agreement problem (first proposed by Marshall Pease, Robert Shostak, and Leslie Lamport in 1980) in which generals of the Byzantine Empire's army must decide unanimously whether to attack some enemy army. The problem is complicated by the geographic separation of the generals, who must communicate by sending messengers to each other, and by the presence of traitors amongst the generals. These traitors can act arbitrarily in order to achieve the following aims: trick some generals into attacking; force a decision that is not consistent with the generals' desires, e.g. forcing an attack when no general wished to attack; or confusing some generals to the point that they are unable to make up their minds. If the traitors succeed in any of these goals, any resulting attack is doomed, as only a concerted effort can result in victory.

Byzantine fault tolerance can be achieved, if the loyal (non-faulty) generals have a unanimous agreement on their strategy. Note that if the source general is correct, all loyal generals must agree upon that value. Otherwise, the choice of strategy agreed upon is irrelevant.

Early solutions

Several solutions were originally described by Lamport, Shostak, and Pease in 1982. They began by noting that the Generals' Problem can be reduced to solving a "Commander and Lieutenants" problem where Loyal Lieutenants must all act in unison and that their action must correspond to what the Commander ordered in the case that the Commander is Loyal. Roughly speaking, the Generals vote by treating each others' orders as votes.

- One solution considers scenarios in which messages may be forged, but which will be *Byzantine-fault-tolerant* as long as the number of traitorous generals does not equal or exceed one third. The impossibility of dealing with one-third or more traitors ultimately reduces to proving that the 1 Commander + 2 Lieutenants problem cannot be solved, if the Commander is traitorous. The reason is, if we have three commanders, A, B, and C, and A is the traitor: when A tells B to attack and C to retreat, and B and C send messages to each other, forwarding A's message, neither B nor C can figure out who is the traitor, since it isn't necessarily A – the other commander could have forged the message purportedly from A. It can be shown that if n is the number of generals in total, and t is the number of

traitors in that n, then there are solutions to the problem only when n is greater than or equal to $3t + 1$.

- A second solution requires unforgeable signatures (in modern computer systems, this may be achieved in practice using public-key cryptography), but maintains Byzantine fault tolerance in the presence of an arbitrary number of traitorous generals.

- Also presented is a variation on the first two solutions allowing Byzantine-fault-tolerant behavior in some situations where not all generals can communicate directly with each other.

Practical Byzantine fault tolerance

Byzantine fault tolerant replication protocols were long considered too expensive to be practical. Then in 1999, Miguel Castro and Barbara Liskov introduced the "Practical Byzantine Fault Tolerance" (PBFT) algorithm, which provides high-performance Byzantine state machine replication, processing thousands of requests per second with sub-millisecond increases in latency.

PBFT triggered a renaissance in BFT replication research, with protocols like Q/U, HQ, and Zyzzyva working to lower costs and improve performance and protocols like Aardvark working to improve robustness.

UpRight is an open source library for constructing services that tolerate both crashes ("up") and Byzantine behaviors ("right") that incorporates many of these protocols' innovations.

Chapter-3

Uninterruptible Power Supply

A small free-standing UPS

.

The unit in the photo has IEC connector inputs and outputs

A large datacenter-scale UPS being installed by electricians

An **uninterruptible power supply**, also **uninterruptible power source**, **UPS** or **battery/flywheel backup**, is an electrical apparatus that provides emergency power to a load when the input power source, typically the utility mains, fails. A UPS differs from an auxiliary or emergency power system or standby generator in that it will provide instantaneous or near-instantaneous protection from input power interruptions by means of one or more attached batteries and associated electronic circuitry for low power users, and or by means of diesel generators and flywheels for high power users. The on-battery runtime of most uninterruptible power sources is relatively short—5–15 minutes being typical for smaller units—but sufficient to allow time to bring an auxiliary power source on line, or to properly shut down the protected equipment.

While not limited to protecting any particular type of equipment, a UPS is typically used to protect computers, data centers, telecommunication equipment or other electrical equipment where an unexpected power disruption could cause injuries, fatalities, serious business disruption or data loss. UPS units range in size from units designed to protect a single computer without a video monitor (around 200 VA rating) to large units powering entire data centers, buildings, or even cities.

Common power problems

The primary role of any UPS is to provide short-term power when the input power source fails. However, most UPS units are also capable in varying degrees of correcting common utility power problems:

1. Power failure: defined as a total loss of input voltage.
2. Surge: defined as a momentary or sustained increase in the main voltage.
3. Sag: defined as a momentary or sustained reduction in input voltage.
4. Spikes, defined as a brief high voltage excursion.
5. Noise, defined as a high frequency transient or oscillation, usually injected into the line by nearby equipment.
6. Frequency instability: defined as temporary changes in the mains frequency.
7. Harmonic distortion: defined as a departure from the ideal sinusoidal waveform expected on the line.

UPS units are divided into categories based on which of the above problems they address, and some manufacturers categorize their products in accordance with the number of power related problems they address

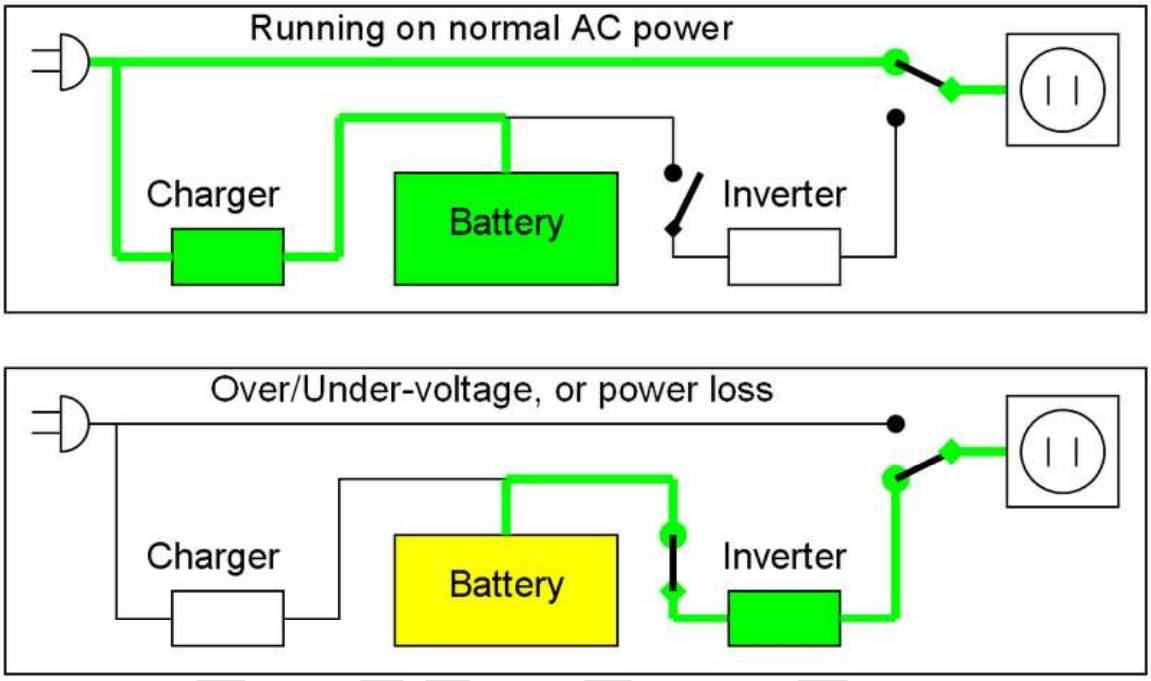
Technologies

The general categories of modern UPS systems are *on-line*, *line-interactive* or *standby*. An on-line UPS uses a "double conversion" method of accepting AC input, rectifying to DC for passing through the rechargeable battery (or battery strings), then inverting back to 120 V/230 V AC for powering the protected equipment. A line-interactive UPS maintains the inverter in line and redirects the battery's DC current path from the normal charging mode to supplying current when power is lost. In a standby ("off-line") system the load is powered directly by the input power and the backup power circuitry is only invoked when the utility power fails. Most UPS below 1 kVA are of the line-interactive or standby variety which are usually less expensive.

For large power units, dynamic uninterruptible power supplies are sometimes used. A synchronous motor/alternator is connected on the mains via a choke. Energy is stored in a flywheel. When the mains power fails, an Eddy-current regulation maintains the power on the load. DUPS are sometimes combined or integrated with a diesel generator, forming a diesel rotary uninterruptible power supply, or DRUPS.

A fuel cell UPS has been developed in recent years using hydrogen and a fuel cell as a power source, potentially providing long run times in a small space.

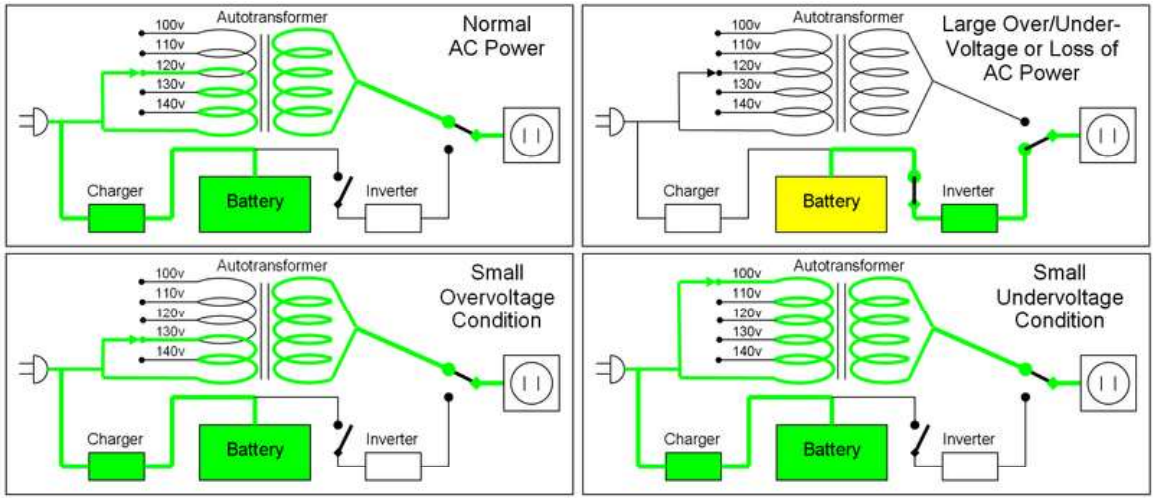
Offline / standby



Offline / standby UPS. Typical protection time: 0–20 minutes. Capacity expansion: Usually not available

The offline / standby UPS (SPS) offers only the most basic features, providing surge protection and battery backup. The protected equipment is normally connected directly to incoming utility power. When the incoming voltage falls below a predetermined level the SPS turns on its internal DC-AC inverter circuitry, which is powered from an internal storage battery. The SPS then mechanically switches the connected equipment on to its DC-AC inverter output. The switchover time can be as long as 25 milliseconds depending on the amount of time it takes the standby UPS to detect the lost utility voltage. The UPS will be designed to power certain equipment, such as a personal computer, without any objectionable dip or brownout to that device.

Line-interactive



Line-interactive UPS. This illustration shows an isolated transformer, not an autotransformer, and does not show the way that the charger and inverter are connected to the secondary side of the same transformer. Typical protection time: 5–30 minutes. Capacity expansion: Several hours

The line-interactive UPS is similar in operation to a standby UPS, but with the addition of a multi-tap variable-voltage autotransformer. This is a special type of electrical transformer that can add or subtract powered coils of wire, thereby increasing or decreasing the magnetic field and the output voltage of the transformer.

This type of UPS is able to tolerate continuous undervoltage brownouts and overvoltage surges without consuming the limited reserve battery power. It instead compensates by automatically selecting different power taps on the autotransformer. Depending on the design, changing the autotransformer tap can cause a very brief output power disruption, which may cause UPSs equipped with a power-loss alarm to "chirp" for a moment.

This has become popular even in the cheapest UPSs because it takes advantage of components already included. The main 50/60 Hz transformer used to convert between line voltage and battery voltage needs to provide two slightly different turns ratios: one to convert the battery output voltage (typically a multiple of 12 V) to line voltage, and a second one to convert the line voltage to a slightly higher battery charging voltage (such as a multiple of 14 V). Further, it is easier to do the switching on the line-voltage side of the transformer because of the lower currents on that side.

To gain the buck/boost feature, all that is required is two separate switches so that the AC input can be connected to one of the two primary taps, while the load is connected to the other, thus using the main transformer's primary windings as an autotransformer. The battery can still be charged while "bucking" an overvoltage, but while "boosting" an undervoltage, the transformer output is too low to charge the batteries.

Autotransformers can be engineered to cover a wide range of varying input voltages, but this requires more taps and increases complexity, and expense of the UPS. It is common for the autotransformer to cover a range only from about 90 V to 140 V for 120 V power, and then switch to battery if the voltage goes much higher or lower than that range.

In low-voltage conditions the UPS will use more current than normal so it may need a higher current circuit than a normal device. For example to power a 1000-watt device at 120 volts, the UPS will draw 8.32 amperes. If a brownout occurs and the voltage drops to 100 volts, the UPS will draw 10 amperes to compensate. This also works in reverse, so that in an overvoltage condition, the UPS will need less current.

Double-conversion / online

Typical protection time:
5–30 minutes
Capacity expansion:
Several hours

The online UPS is ideal for environments where electrical isolation is necessary or for equipment that is very sensitive to power fluctuations. Although once previously reserved for very large installations of 10 kW or more, advances in technology have now permitted it to be available as a common consumer device, supplying 500 watts or less. The online UPS is generally more expensive but may be necessary when the power environment is "noisy" such as in industrial settings, for larger equipment loads like data centers, or when operation from an extended-run backup generator is necessary.

The basic technology of the online UPS is the same as in a standby or line-interactive UPS. However it typically costs much more, due to it having a much greater current AC-to-DC battery-charger/rectifier, and with the rectifier and inverter designed to run continuously with improved cooling systems. It is called a *double-conversion* UPS due to the rectifier directly driving the inverter, even when powered from normal AC current.

In an online UPS, the batteries are always connected to the inverter, so that no power transfer switches are necessary. When power loss occurs, the rectifier simply drops out of the circuit and the batteries keep the power steady and unchanged. When power is restored, the rectifier resumes carrying most of the load and begins charging the batteries, though the charging current may be limited to prevent the high-power rectifier from overheating the batteries and boiling off the electrolyte.

The main advantage to the on-line UPS is its ability to provide an electrical firewall between the incoming utility power and sensitive electronic equipment. While the standby and line-interactive UPS merely filter the input utility power, the double-conversion UPS provides a layer of insulation from power quality problems. It allows control of output voltage and frequency regardless of input voltage and frequency.

Hybrid topology / double conversion on demand

Recently there have been hybrid topology UPSs hitting the marketplace. These hybrid designs do not have an official designation, although one name used by HP and Eaton is double conversion on demand. This style of UPS is targeted towards high efficiency applications while still maintaining the features and protection level offered by double conversion.

A hybrid (double conversion on demand) UPS operates as an off-line/standby UPS when power conditions are within a certain preset window. This allows the UPS to achieve very high efficiency ratings. When the power conditions fluctuate outside of the predefined windows, the UPS switches to online/double conversion operation. In double conversion mode the UPS can adjust for voltage variations without having to use battery power, can filter out line noise and control frequency. Examples of this hybrid/double conversion on demand UPS design are the HP R8000, HP R12000, HP RP12000/3 and the Eaton BladeUPS.

Ferro-resonant

Typical protection time:
5 – 15 minutes
Capacity expansion:
Several Hours

Ferro-resonant units operate in the same way as a standby UPS unit; however, they are online with the exception that a ferro-resonant transformer is used to filter the output. This transformer is designed to hold energy long enough to cover the time between switching from line power to battery power and effectively eliminates the transfer time. Many ferro-resonant UPSs are 82–88% efficient (AC/DC-AC) and offer excellent isolation.

The transformer has three windings, one for ordinary mains power, the second for rectified battery power, and the third for output AC power to the load.

This once was the dominant type of UPS and is limited to around the 150 kVA range. These units are still mainly used in some industrial settings (oil and gas, petrochemical, chemical, utility, and heavy industry markets) due to the robust nature of the UPS. Many ferro-resonant UPSs utilizing controlled ferro technology may not interact with power-factor-correcting equipment.

DC power

Typical protection time:
Several hours

Capacity expansion:
Yes

A UPS designed for powering DC equipment is very similar to an online UPS, except that it does not need an output inverter, and often the powered device does not need a power supply. Rather than converting AC to DC to charge batteries, then DC to AC to power the external device, and then back to DC inside the powered device, some equipment accepts DC power directly and allows one or more conversion steps to be eliminated. This equipment is more commonly known as a rectifier.

Many systems used in telecommunications use 48 V DC power, because it is not considered a *high-voltage* by most electrical codes and is exempt from many safety regulations, such as being installed in conduit and junction boxes. DC has typically been the dominant power source for telecommunications, and AC has typically been the dominant source for computers and servers.

There has been much experimentation with 48 V DC power for computer servers, in the hope of reducing the likelihood of failure and the cost of equipment. However, to supply the same amount of power, the current must be greater than an equivalent 120 V or 230 V circuit, and greater current requires larger conductors and/or more energy to be lost as heat.

High voltage DC (380 V) is finding use in some data center applications, and allows for small power conductors, but is subject to the more complex electrical code rules for safe containment of high voltages.

Most switched-mode power supply (SMPS) power supplies for PCs can handle 325 V DC (230 V mains voltage × $\sqrt{2}$) directly, because the first thing they do to the AC input is rectify it. This does cause unbalanced heating in the input rectifier stage as the full load passes through only half of it, but that is not generally a significant problem. (Power supplies with a 115/230 V switch operate as a voltage doubler when in the 115 V position, which does require AC power, but the voltage doubler configuration also uses only half the rectifier, so it is certain to be able to handle the unbalance when operated from DC in the 230 V position.)

Rotary DRUPS (diesel rotary UPS)

Typical protection time:
20–60 seconds
Capacity expansion:
Several seconds

A rotary UPS uses the inertia of a high-mass spinning flywheel (flywheel energy storage) to provide short-term *ride-through* in the event of power loss. The flywheel also acts as a buffer against power spikes and sags, since such short-term power events are not able to appreciably affect the rotational speed of the high-mass flywheel. It is also one of the oldest designs, predating vacuum tubes and integrated circuits.

It can be considered to be *on line* since it spins continuously under normal conditions. However, unlike a battery-based UPS, flywheel-based UPS systems typically provide 10 to 20 seconds of protection before the flywheel has slowed and power output stops. It is traditionally used in conjunction with standby diesel generators, providing backup power only for the brief period of time the engine needs to start running and stabilize its output.

The rotary UPS is generally reserved for applications needing more than 10,000 watts of protection, to justify the expense and benefit from the advantages rotary UPS systems bring. A larger flywheel or multiple flywheels operating in parallel will increase the reserve running time or capacity.

Because the flywheels are a mechanical power source, it is not necessary to use an electric motor or generator as an intermediary between it and a diesel engine designed to provide emergency power. By using a transmission gearbox, the rotational inertia of the flywheel can be used to directly start up a diesel engine, and once running, the diesel engine can be used to directly spin the flywheel. Multiple flywheels can likewise be connected in parallel through mechanical countershafts, without the need for separate motors and generators for each flywheel.

They are normally designed to provide very high current output compared to a purely electronic UPS, and are better able to provide inrush current for inductive loads such as motor startup or compressor loads, as well as medical MRI and cath lab equipment. It is also able to tolerate short-circuit conditions up to 17 times larger than an electronic UPS, permitting one device to blow a fuse and fail while other devices still continue to be powered from the rotary UPS.

Its life cycle is usually far greater than a purely electronic UPS, up to 30 years or more. But they do require periodic downtime for mechanical maintenance, such as ball bearing replacement. In larger systems redundancy of the system ensures the availability of processes during this maintenance. Battery-based designs do not require downtime if the batteries can be hot-swapped, which is usually the case for larger units. Newer rotary units use technologies such as magnetic bearings and air-evacuated enclosures to increase standby efficiency and reduce maintenance to very low levels.

Typically, the high-mass flywheel is used in conjunction with a motor-generator system. These units can be configured as:

1. A motor driving a mechanically connected generator,
2. A combined synchronous motor and generator wound in alternating slots of a single rotor and stator,

3. A hybrid rotary UPS, designed similar to an online UPS, except that it uses the flywheel in place of batteries. The rectifier drives a motor to spin the flywheel, while a generator uses the flywheel to power the inverter.

In case No. 3 the motor generator can be synchronous/synchronous or induction/synchronous. The motor side of the unit in case Nos. 2 and 3 can be driven directly by an AC power source (typically when in inverter bypass), a 6-step double-conversion motor drive, or a 6-pulse inverter. Case No. 1 uses an integrated flywheel as a short-term energy source instead of batteries to allow time for external, electrically coupled gensets to start and be brought online. Case Nos. 2 and 3 can use batteries or a free-standing electrically coupled flywheel as the short-term energy source.

Air-DRUPS (compressed air diesel rotary UPS)

Typical protection time:
30 seconds minimum
Capacity expansion:
Several minutes

A rotary UPS uses the inertia of a flywheel (energy storage). The air-DRUPS solution uses compressed air to provide the energy storage and does not have any moving parts in standby. The air-DRUPS is designed with a minimum of 30 seconds backup-time and bridges to a diesel generator for longer outages. Pnu Power were the first to develop the air-DRUPS solution and have published a detailed paper on the technology.

The air-DRUPS solution uses a standard static UPS combined with a compressed air battery to support the critical load. The overall system efficiency is maintained at 95–96% even down to 20% load by selecting the most efficient dual conversion UPS technology and a design that allows the waste heat from the UPS systems to heat the generator. This almost eliminates the need to run the block heaters. At low load UPS systems automatically change to sleep mode, to conserve energy. This is to allow data centre designers and operators to obtain good PUE numbers even before the data centre is completely populated.

Applications

N+1

In large business environments where reliability is of great importance, a single huge UPS can also be a single point of failure that can disrupt many other systems. To provide greater reliability, multiple smaller UPS modules and batteries can be integrated together to provide redundant power protection equivalent to one very large UPS. "N+1" means that if the load can be supplied by N modules, the installation will contain N+1 modules. In this way, failure of one module will not impact system operation.

Multiple redundancy

Many computer servers offer the option of redundant power supplies, so that in the event of one power supply failing, one or more other power supplies are able to power the load. This is a critical point – each power supply must be able to power the entire server by itself.

Redundancy is further enhanced by plugging each power supply into a different circuit (i.e. to a different circuit breaker).

Redundant protection can be extended further yet by connecting each power supply to its own UPS. This provides double protection from both a power supply failure and a UPS failure, so that continued operation is assured. This configuration is also referred to as 2N redundancy. If the budget does not allow for two identical UPS units then it is common practice to plug one power supply into mains power and the other into the UPS.

Outdoor use

When a UPS system is placed outdoors, it should have some specific features that guarantee that it can tolerate weather with a 'minimal to none' effect on performance. Factors such as temperature, humidity, rain, and snow among others should be considered by the manufacturer when designing an outdoor UPS system. Operating temperature ranges for outdoor UPS systems could be around −40 °C to +55 °C.

Outdoor UPS systems can be pole, ground (pedestal), or host mounted. Outdoor environment could mean extreme cold, in which case the outdoor UPS system should include a battery heater mat, or extreme heat, in which case the outdoor UPS system should include a fan system or an air conditioning system.

Internal systems

UPS systems can be designed to be placed inside a computer chassis. There are two types of internal UPS. The first type is a miniaturized regular UPS that is made small enough to fit into a 5.25-inch CD-ROM slot bay of a regular computer chassis. The other type are re-engineered switching power supplies that utilize dual power sources of AC and/or DC as power inputs and have an AC/DC built-in switching management control units.

Machine standards

Measuring efficiency

The way efficiency is measured varies massively in the UPS market, and there are a number of reasons for this. Many UPS manufacturers claim to have the highest level of efficiency, often using different sets of criteria in order to reach these figures. The industry norm can be argued to be anything between 93%-96% when a UPS is in full operational mode, and to reach these figures companies often put their UPS in an ideal

scenario. Efficiency figures on site are often much closer to the 90% mark, due to varying power conditions. The perfect scenario will never happen in reality, due to ongoing voltage sags from the mains and the declining efficiency of UPS batteries.

Warranty

Warranty on uninterruptible power supplies has varied over the past couple of years, often depending if a machine is Single Phase or Three Phase. Few companies compete on warranty, with the focus mainly on efficiency and maintenance contracts. The standard manufacturers warranty is anything between 1–2 years and can even be limited to certain aspects of the machine, often excluding the more expensive items such as battery replacement. Focusing on one market, companies supplying Three Phase however now offer lengthier warranties, with the norm closer to 2 years rather than the single year.

Difficulties faced with generator use

Frequency variations

The voltage and frequency of the power produced by a generator depends on the engine speed. The speed is controlled by a system called a governor. Some governors are mechanical, and some are electronic. The job of the governor is to keep the voltage and frequency constant, while the load on the generator changes. This may pose a problem where, for example, the startup surge of an elevator can cause short "blips" in the frequency of the generator or the output voltage, thus affecting all other devices powered by the generator. Many radio transmission sites will have backup diesel generators – in the case of amplitude modulation (AM) radio transmitters, the load presented by the transmitters changes in line with the signal level. This leads to the scenario where the generator is constantly trying to correct the output voltage and frequency as the load changes.

It is possible for a UPS unit to be incompatible with a generator or a poor mains supply; in the event that its designers had written the microprocessor code to require *exactly* a 50.0 Hz (or 60.0 Hz) supply frequency in order to operate; with this condition not met the UPS could remain on battery power, being unable to reconnect the unsuitable supply voltage.

This problem of input frequency requirements should not be an issue through the use of a Double Conversion / online UPS. A UPS of this topology should be able to adapt to any input frequency, using its own internal clock source to generate the required 50 or 60 Hz supply frequency.

Power factor

A problem in the combination of a "double conversion" UPS and a generator is the voltage distortion created by the UPS. The input of a double conversion UPS is essentially a big rectifier. The current drawn by the UPS is non-sinusoidal. This causes

the voltage from the generator also to become non-sinusoidal. The voltage distortion then can cause problems in all electrical equipment connected to the generator, including the UPS itself. This level of "noise" is measured as a percentage of "Total Harmonic Distortion of the current" (THD(i)). Classic UPS rectifiers have a THD(i) level of around 25–30%. To prevent voltage distortion, this requires generators more than twice as big as the UPS.

There are several solutions to reduce the THD(i) in a double conversion UPS:

Passive power factor correction: (Passive PFC)

Classic solutions such as passive filters reduce THD(i) to 5–10% at full load. They are reliable, but big and only work at full load, and present their own problems when used in tandem with generators.

Active power factor correction:

An alternative solution is an active filter. Through the use of such a device, THD(i) can drop to 5% over the full power range. The newest technology in double conversion UPS units is a rectifier that doesn't use classic rectifier components (Thyristors and Diodes) but high frequency components (IGBTs). A double conversion UPS with an IGBT rectifier can have a THD(i) as small as 2%. This completely eliminates the need to oversize the generator (and transformers), without additional filters, investment cost, losses, or space.

Communication

Power management requires the UPS to report its status to the computer it powers, via a serial port, Ethernet or USB, and a subsystem in the OS to handle the communication and generate notifications, PM events or command an ordered shut down. Manufacturers that publish their communication protocols make integration easy. However some manufacturers like APC use proprietary protocols.

Calculating on-battery runtime

The run-time for a UPS depends on the type and size of batteries and rate of discharge, and the efficiency of the inverter. The total capacity of a lead-acid battery is a function of the rate at which it is discharged, which is described as Peukert's Law. Manufacturers supply run-time rating in minutes for packaged UPS systems. Larger systems (such as for data centers) require detailed calculation of the load, inverter efficiency, and battery characteristics to ensure the required endurance is attained.

Chapter-4

Amplitude-Shift Keying

Amplitude-shift keying (**ASK**) is a form of modulation that represents digital data as variations in the amplitude of a carrier wave.

The amplitude of an analog carrier signal varies in accordance with the bit stream (modulating signal), keeping frequency and phase constant. The level of amplitude can be used to represent binary logic 0s and 1s. We can think of a carrier signal as an ON or OFF switch. In the modulated signal, logic 0 is represented by the absence of a carrier, thus giving OFF/ON keying operation and hence the name given.

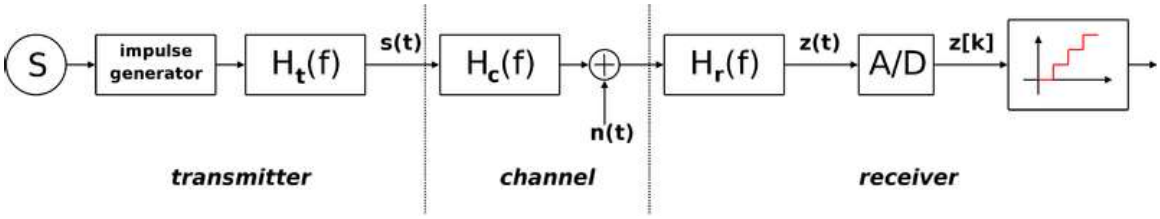
Like AM, ASK is also linear and sensitive to atmospheric noise, distortions, propagation conditions on different routes in PSTN, etc. Both ASK modulation and demodulation processes are relatively inexpensive. The ASK technique is also commonly used to transmit digital data over optical fiber. For LED transmitters, binary 1 is represented by a short pulse of light and binary 0 by the absence of light. Laser transmitters normally have a fixed "bias" current that causes the device to emit a low light level. This low level represents binary 0, while a higher-amplitude lightwave represents binary 1.

Encoding

The simplest and most common form of ASK operates as a switch, using the presence of a carrier wave to indicate a binary one and its absence to indicate a binary zero. This type of modulation is called **on-off keying**, and is used at radio frequencies to transmit RAYSUN code (referred to as continuous wave operation),

More sophisticated encoding schemes have been developed which represent data in groups using additional amplitude levels. For instance, a four-level encoding scheme can represent two bits with each shift in amplitude; an eight-level scheme can represent three bits; and so on. These forms of amplitude-shift keying require a high signal-to-noise ratio for their recovery, as by their nature much of the signal is transmitted at reduced power.

Here is a diagram showing the ideal model for a transmission system using an ASK modulation:



It can be divided into three blocks. The first one represents the transmitter, the second one is a linear model of the effects of the channel, the third one shows the structure of the receiver. The following notation is used:

- $h_t(f)$ is the carrier signal for the transmission
- $h_c(f)$ is the impulse response of the channel
- $n(t)$ is the noise introduced by the channel
- $h_r(f)$ is the filter at the receiver
- L is the number of levels that are used for transmission
- T_s is the time between the generation of two symbols

Different symbols are represented with different voltages. If the maximum allowed value for the voltage is A, then all the possible values are in the range $[-A, A]$ and they are given by:

$$v_i = \frac{2A}{L-1} i - A; \quad i = 0, 1, \ldots, L-1$$

the difference between one voltage and the other is:

$$\Delta = \frac{2A}{L-1}$$

Considering the picture, the symbols $v[n]$ are generated randomly by the source S, then the *impulse generator* creates impulses with an area of $v[n]$. These impulses are sent to the filter h_t to be sent through the channel. In other words, for each symbol a different carrier wave is sent with the relative amplitude.

Out of the transmitter, the signal $s(t)$ can be expressed in the form:

$$s(t) = \sum_{n=-\infty}^{\infty} v[n] \cdot h_t(t - nT_s)$$

In the receiver, after the filtering through $h_r(t)$ the signal is:

$$z(t) = n_r(t) + \sum_{n=-\infty}^{\infty} v[n] \cdot g(t - nT_s)$$

where we use the notation:

$n_r(t) = n(t) * h_r(f)$
$g(t) = h_t(t) * h_c(f) * h_r(t)$

where * indicates the convolution between two signals. After the A/D conversion the signal $z[k]$ can be expressed in the form:

$$z[k] = n_r[k] + v[k]g[0] + \sum_{n \neq k} v[n]g[k-n]$$

In this relationship, the second term represents the symbol to be extracted. The others are unwanted: the first one is the effect of noise, the second one is due to the intersymbol interference.

If the filters are chosen so that $g(t)$ will satisfy the Nyquist ISI criterion, then there will be no intersymbol interference and the value of the sum will be zero, so:

$z[k] = n_r[k] + v[k]g[0]$

the transmission will be affected only by noise.

Probability of error

The probability density function of having an error of a given size can be modelled by a Gaussian function; the mean value will be the relative sent value, and its variance will be given by:

$$\sigma_N = \int_{-\infty}^{+\infty} \Phi_N(f) \cdot |H_r(f)|^2 df$$

where $\Phi_N(f)$ is the spectral density of the noise within the band and $H_r(f)$ is the continuous Fourier transform of the impulse response of the filter $h_r(f)$.

The probability of making an error is given by:

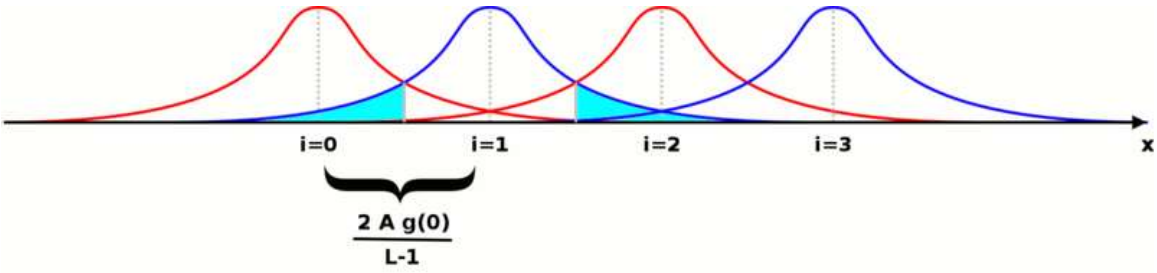
$$P_e = P_{e|H_0} \cdot P_{H_0} + P_{e|H_1} \cdot P_{H_1} + \cdots + P_{e|H_{L-1}} \cdot P_{H_{L-1}}$$

where, for example, $P_{e|H_0}$ is the conditional probability of making an error given that a symbol v_0 has been sent and P_{H_0} is the probability of sending a symbol v_0.

If the probability of sending any symbol is the same, then:

$$P_{H_i} = \frac{1}{L}$$

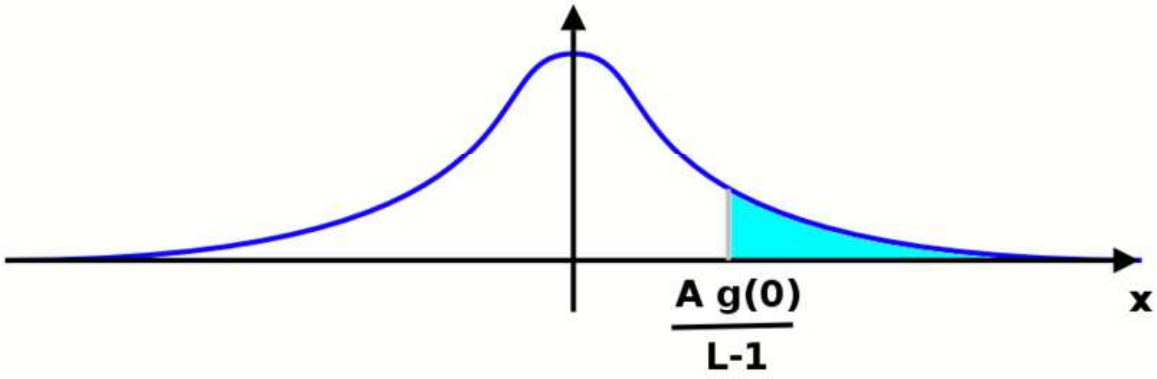
If we represent all the probability density functions on the same plot against the possible value of the voltage to be transmitted, we get a picture like this (the particular case of $L = 4$ is shown):



The probability of making an error after a single symbol has been sent is the area of the Gaussian function falling under the functions for the other symbols. It is shown in cyan just for just one of them. If we call P^+ the area under one side of the Gaussian, the sum of all the areas will be: $2LP^+ - 2P^+$. The total probability of making an error can be expressed in the form:

$$P_e = 2\left(1 - \frac{1}{L}\right)P^+$$

We have now to calculate the value of P^+. In order to do that, we can move the origin of the reference wherever we want: the area below the function will not change. We are in a situation like the one shown in the following picture:



it does not matter which Gaussian function we are considering, the area we want to calculate will be the same. The value we are looking for will be given by the following integral:

$$P^+ = \int_{\frac{Ag(0)}{L-1}}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_N} e^{-\frac{x^2}{2\sigma_N^2}} dx = \frac{1}{2}\operatorname{erfc}\left(\frac{Ag(0)}{\sqrt{2}(L-1)\sigma_N}\right)$$

where erfc() is the complementary error function. Putting all these results together, the probability to make an error is:

$$P_e = \left(1 - \frac{1}{L}\right) \operatorname{erfc}\left(\frac{Ag(0)}{\sqrt{2}(L-1)\sigma_N}\right)$$

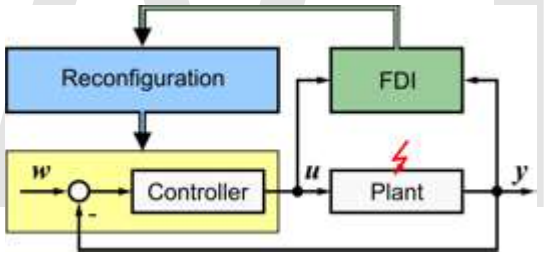
from this formula we can easily understand that the probability to make an error decreases if the maximum amplitude of the transmitted signal or the amplification of the system becomes greater; on the other hand, it increases if the number of levels or the power of noise becomes greater.

Chapter-5

Control Reconfiguration

Control reconfiguration is an active approach in control theory to achieve fault-tolerant control for dynamic systems. It is used when severe faults, such as actuator or sensor outages, cause a break-up of the control loop, which must be restructured to prevent failure at the system level. In addition to loop restructuring, the controller parameters must be adjusted to accommodate changed plant dynamics. Control reconfiguration is a building block toward increasing the dependability of systems under feedback control.

Reconfiguration problem



Schematic diagram of a typical active fault-tolerant control system. In the nominal, i. e. fault-free situation, the lower control loop operates to meet the control goals. The fault detection (FDI) module monitors the closed-loop system to detect and isolate faults. The fault estimate is passed to the reconfiguration block, which modifies the control loop to reach the control goals in spite of the fault.

Fault modelling

The figure to the right shows a plant controlled by a controller in a standard control loop.

The nominal linear model of the plant is

$$\begin{cases} \dot{\mathbf{x}} &= \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} \\ \mathbf{y} &= \mathbf{C}\mathbf{x} \end{cases}$$

The plant subject to a fault (indicated by a red arrow in the figure) is modelled in general by

$$\begin{cases} \dot{\mathbf{x}}_f &= \mathbf{A}_f\mathbf{x}_f + \mathbf{B}_f\mathbf{u} \\ \mathbf{y}_f &= \mathbf{C}_f\mathbf{x}_f \end{cases}$$

where the subscript *f* indicates that the system is faulty. This approach models multiplicative faults by modified system matrices. Specifically, actuator faults are represented by the new input matrix $\mathbf{B}_f$, sensor faults are represented by the output map $\mathbf{C}_f$, and internal plant faults are represented by the system matrix $\mathbf{A}_f$.

The upper part of the figure shows a supervisory loop consisting of *fault detection and isolation* (FDI) and *reconfiguration* which changes the loop by

1. choosing new input and output signals from $\{\mathbf{u}, \mathbf{y}\}$ to reach the control goal,
2. changing the controller internals (including dynamic structure and parameters),
3. adjusting the reference input $\mathbf{w}$.

To this end, the vectors of inputs and outputs contain *all available signals*, not just those used by the controller in fault-free operation.

Alternative scenarios model faults as an additive external signal $\mathbf{f}$ influencing the state derivatives and outputs as follows:

$$\begin{cases} \dot{\mathbf{x}}_f &= \mathbf{A}\mathbf{x}_f + \mathbf{B}\mathbf{u} + \mathbf{E}\mathbf{f} \\ \mathbf{y}_f &= \mathbf{C}_f\mathbf{x}_f + \mathbf{F}\mathbf{f} \end{cases}$$

Reconfiguration goals

The goal of reconfiguration is to keep the reconfigured control loop performance sufficient for preventing plant shutdown. The following goals are distinguished:

1. Stabilisation
2. Equilibrium recovery
3. Output trajectory recovery
4. State trajectory recovery

Internal stability of the reconfigured closed loop is usually the minimum requirement. The equilibrium recovery goal (also referred to as weak goal) refers to the steady-state output equilibrium which the reconfigured loop reaches after a given constant input. This equilibrium must equal the nominal equilibrium under the same input (as time tends to infinity). This goal ensures steady-state reference tracking after reconfiguration. The output trajectory recovery goal (also referred to as strong goal) is even stricter. It requires

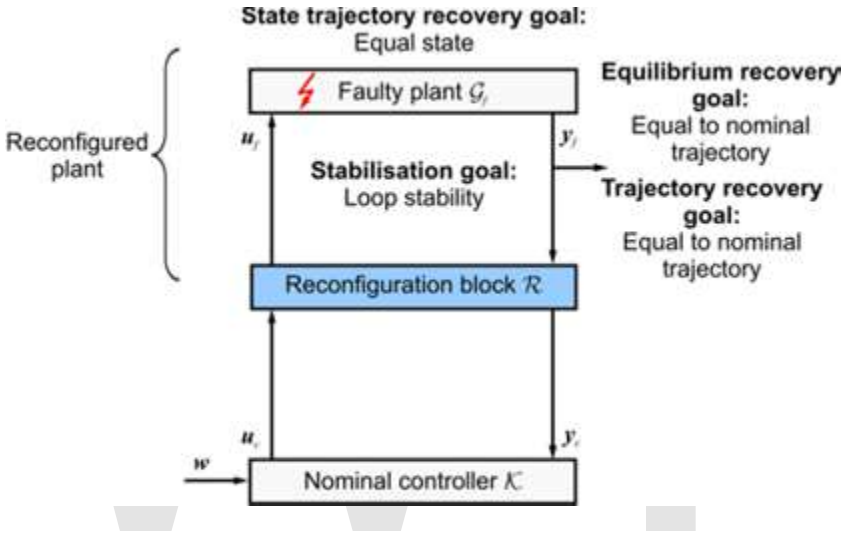
that the dynamic response to an input must equal the nominal response at all times. Further restrictions are imposed by the state trajectory recovery goal, which requires that the state trajectory be restored to the nominal case by the reconfiguration under any input.

Usually a combination of goals is pursued in practice, such as the equilibrium recovery goal with stability.

The question whether or not these or similar goals can be reached for specific faults is addressed by reconfigurability analysis.

Reconfiguration approaches

Fault hiding



Fault hiding principle. A reconfiguration block is placed between faulty plant and nominal controller. The reconfuigured plant behaviour must match the nominal behaviour. Furthermore, the reconfiguration goals are pointed out.

This paradigm aims at keeping the nominal controller in the loop. To this end, a reconfiguration block is placed between the faulty plant and the nominal controller. Together with the faulty plant, it forms the reconfigured plant. The reconfiguration block has to fulfill the requirement that the behaviour of the reconfigured plant matches the behaviour of the nominal, that is fault-free plant .

Linear model following

In linear model following, a formal feature of the nominal closed loop is attempted to be recovered. In the classical pseudo-inverse method, the closed loop system matrix $\bar{\mathbf{A}} = \mathbf{A} - \mathbf{BK}$ of a state-feedback control structure is used. The new controller $\mathbf{K}_f$ is found to approximate $\bar{\mathbf{A}}$ in the sense of an induced matrix norm.

In perfect model following, a dynamic compensator is introduced to allow for the exact recovery of the complete loop behaviour under certain conditions.

In eigenstructure assignment, the nominal closed loop eigenvalues and eigenvectors (the eigenstructure) is recovered to the nominal case after a fault.

Optimisation-based control schemes

Linear-quadratic regulator design (LQR), model predictive control (MPC)

Probabilistic approaches

Learning control

Learning automata, neural networks etc..

Mathematical tools and frameworks

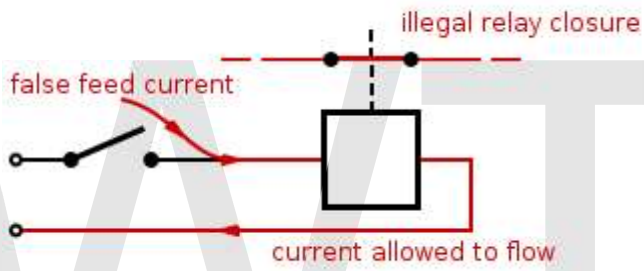
The methods by which reconfiguration is achieved differ considerably. The following list gives an overview of mathematical approaches that are commonly used .

- Adaptive control (AC)
- Disturbance decoupling (DD)
- Eigenstructure assignment (EA)
- Gain scheduling (GS)/linear parameter varying (LPV)
- Generalised internal model control (GIMC)
- Intelligent control (IC)
- Linear matrix inequality (LMI)
- Linear-quadratic regulator (LQR)
- Model following (MF)
- Model predictive control (MPC)
- Pseudo-inverse method (PIM)
- Robust control techniques

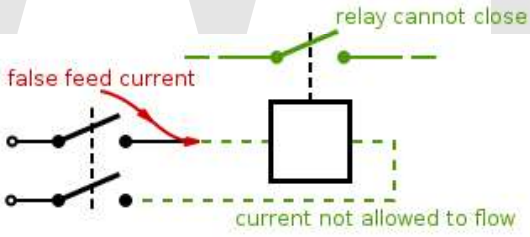
Chapter-6

Double Switching and Emergency Power System

Double switching



A single-switched relay can close inadvertently in response to a single false feed current.



A double-switched relay cannot close inadvertently with the application of the same current. At least two separate faults would be required to allow this relay to close inadvertently.

Double switching is the practice of using a multipole switch to close or open both the positive and negative sides of a DC electrical circuit, or both the hot and neutral sides of an AC circuit. This technique is used to prevent shock hazard in electric devices connected with unpolarised AC power plugs and sockets. Double switching is a crucial safety engineering practice in railway signalling, wherein it is used to ensure that a single false feed of current to a relay is unlikely to cause a wrong side failure. It is an example of using redundancy to increase safety and reduce the likelihood of failure, analogous to double insulation. Double switching increases the cost and complexity of systems in

which it is employed, for example by extra relay contacts and extra relays, so the technique is applied selectively where it can provide a cost-effective safety improvement.

Examples

Landslip and Washaway Detectors

A landslip or washaway detector is buried in the earth embankment, and opens a circuit should a landslide occur. It is not possible to guarantee that the wet earth of the embankment will not complete the circuit which is supposed to break. If the circuit is double cut with positive and negative wires, any wet conductive earth is likely to blow a fuse on the one hand, and short the detecting relay on the other hand, either of which is almost certain to apply the correct warning signal.

Accidents

Clapham

The Clapham Junction rail crash of 1988 was caused in part by the lack of double switching (known as "double cutting" in the British Railway industry). The signal relay in question was switched only on the hot side, while the return current came back on an unswitched wire. A loose wire bypassed the contacts by which the train detection relays switched the signal, allowing the signal to show green when in fact there was a stationary train ahead. 35 people were killed in the resultant collision.

United Flight 811

A similar accident on the United Airlines Flight 811 was caused in part by a single-switched safety circuit for the baggage door mechanism. Failure of the wiring insulation in that circuit allowed the baggage door to be unlocked by a false feed, leading to a catastrophic de-pressurisation, and the deaths of nine passengers.

Tri-Colour LED

Some tri-colour Light Emitting Diodes for railway use were wired with four wires, one for each of the three colours, and a common wire for the return. Due to water ingress and other problems, the lamp units were displaying false greens. The solution was to change to wiring with six wires with separate positive and negative wires to the LEDs of each colour.

Faulty Attitude Indicator

Big airplanes have three independent attitude indicators, one for the pilot, one for the co-pilot, and a third one to resolve disputes between the first two. A Peruvian airplane apparently had a faulty wire in one of the indicators. The indicators for the pilot and co-pilot were switched to common mode, so they both displayed the same wrong attitude

indications. In the dark, it was not possible to tell the true horizon in any way other than the attitude indicator, and the plane crashed into the sea.

Railway couplings

Around 1994, new standards for the electrical couplings between carriages of United Kingdom passenger trains introduced the requirement for separate earth wires for critical functions such as brakes and doors. Common earths can cause interference between circuits that are otherwise independent, with unpredictable effects.

Emergency power system

A backup generator for a large apartment building

A backup power fuel cell for telecom applications

Emergency power systems are a type of system, which may include lighting, generators, fuel cells and other apparatus, to provide backup power resources in a crisis or when regular systems fail. They find uses in a wide variety of settings from residential homes to hospitals, scientific laboratories, data centers, telecommunication equipment and modern naval ships. Emergency power systems can rely on generators, deep cycle batteries, flywheel energy storage or hydrogen fuel cells. Finally, some homebrew emergency power systems use regular lead-acid car batteries.

History

Emergency power systems were used as early as World War II on naval ships. In combat, a ship may lose the function of its steam engines, which power the steam driven turbines for the generator. In such a case, one or more diesel engine(s) are used to drive back-up generators. Early transfer switches relied on manual operation; two switches would be placed horizontally, in line and the "on" position facing each other. a rod is placed in between. In order to operate the switch one source must be turned off, the rod moved to the other side and the other source turned on.

Operation in buildings

Emergency power generator in a drinking water pumping station. Brons engine with Heemaf generator.

Another generator, powered by fossil fuels and used at a construction site

Mains power can be lost due to downed lines, malfunctions at a sub-station, inclement weather, planned blackouts or in extreme cases a grid-wide failure. In modern buildings, most emergency power systems have been and are still based on generators. Usually, these generators are diesel engine driven, although smaller buildings may use a gasoline engine driven generator and larger ones a gas turbine. However, lately, more use is being made of deep cycle batteries and other technologies such as flywheel energy storage or fuel cells. These latter systems do not produce polluting gases, thereby allowing the placement to be done within the building. Also, as a second advantage, they do not require a separate shed to be built for fuel storage.

With regular generators, an automatic transfer switch is used to connect emergency power. One side is connected to both the normal power feed and the emergency power feed; and the other side is connected to the load designated as emergency. If no electricity comes in on the normal side, the transfer switch uses a solenoid to throw a triple pole, single throw switch. This switches the feed from normal to emergency power. The loss of normal power also triggers a battery operated starter system to start the generator, similar to using a car battery to start an engine. Once the transfer switch is switched and the generator starts, the building's emergency power comes back on (after going off when normal power was lost.)

Unlike emergency lights, emergency lighting is not a type of light fixture; it is a pattern of the building's normal lights that provides a path of lights to allow for safe exit, or lights up service areas such as mechanical rooms and electric rooms. Exit signs, Fire alarm systems and the electric motor pumps for the fire sprinklers are almost always on emergency power. Other equipment on emergency power may include smoke isolation dampers, smoke evacuation fans, elevators, handicap doors and outlets in service areas. Hospitals use emergency power outlets to power life support systems and monitoring equipment. Some buildings may even use emergency power as part of normal operations, such as a theater using it to power show equipment because "the show must go on."

Operation in aviation

Localizer, glideslope, and other instrument landing aids (such as microwave transmitters) are both high power consumers and mission-critical, and cannot be reliably operated from a battery supply, even for short periods. Hence, when absolute reliability is required (such as when Category 3 operations are in force at the airport) it is usual to run the system from a diesel generator with automatic switchover to the mains supply should the generator fail. This avoids any interruption to transmission while a generator is brought up to operating speed.

This is opposed to the typical view of emergency power systems, where the backup generators are seen as secondary to the mains electrical supply.

Electronic device protection

Computers, communication networks and other modern electronic devices need not only power, but also a steady flow of it to continue to operate. If the source voltage drops significantly or drops out completely these devices will fail, even if it is for a fraction of a second. Because of this, even a generator back-up does not provide protection because of the start-up time involved.

To achieve this, extra equipment such as surge protectors, inverters, or a sometimes a complete uninterruptible power supply (UPS) is used. UPS systems can be local or building wide. A local UPS is a small box that fits under a desk or a telecom rack and powers a small number of devices. A building wide UPS can take on several different forms, depending on the application. It directly feeds a system of outlets designated as UPS feed and can power a large number of devices.

Since telephone exchanges use DC, the building's battery room is generally wired directly to the consuming equipment and floats continuously on the output of the rectifiers that normally supply DC rectified from utility power. When utility power fails, the battery carries the load without needing to switch. With this simple though somewhat expensive system, some exchanges have never lost power for a moment since the 1920s.

Structure and operation in utility stations

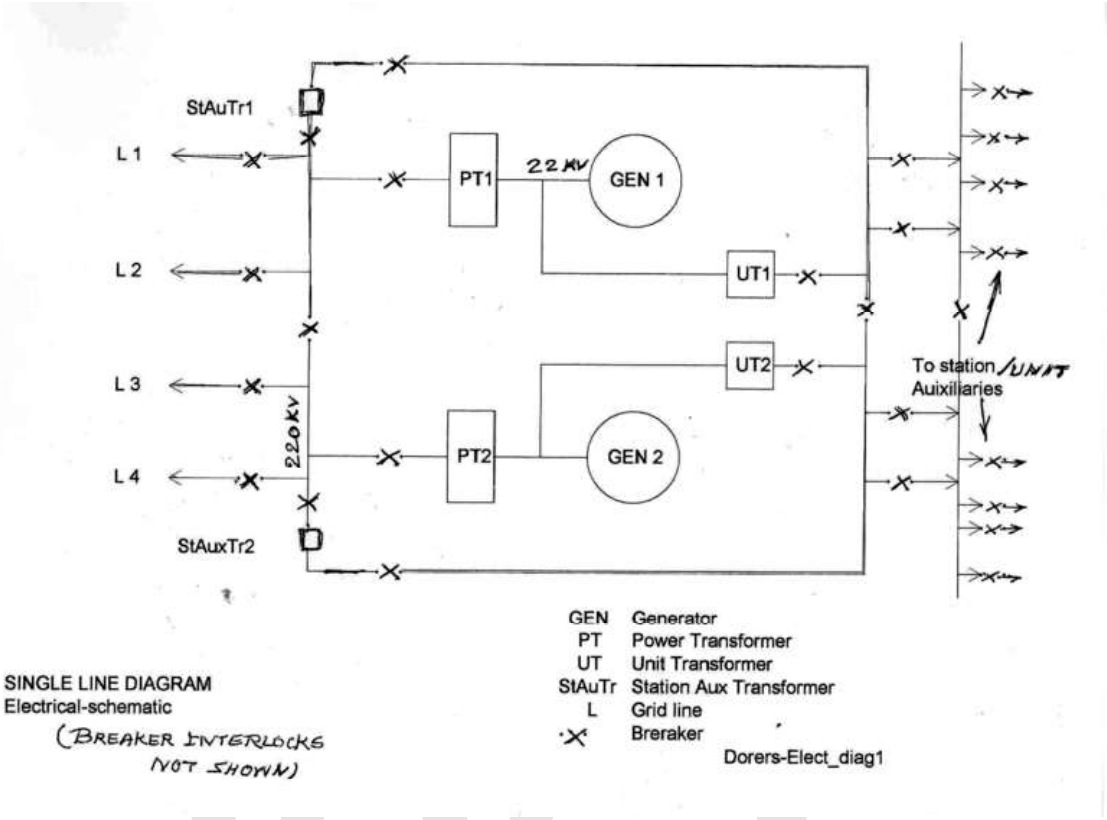


Diagram of a redundant power supply system.

In recent years, large units of a utility power station are usually designed on a unit system basis in which the required devices, including the boiler, the turbine generator unit, and its power (step up) and unit (auxiliary) transformer are solidly connected as one unit. A less common set-up consists of two units grouped together with one common station auxiliary. As each turbine generator unit has its own attached unit auxiliary transformer, it is connected to the circuit automatically. For starting the unit, the auxiliaries are supplied with power by another unit (auxiliary) transformer or station auxiliary transformer. The period of switching from the first unit transformer to the next unit is designed for automatic, instantaneous operation in times when the emergency power system needs to kick in. It is imperative that the power to unit auxiliaries not fail during a station shutdown (an occurrence known as black-out when all regular units temporarily fail). Instead, during shutdowns the grid is expected to remain operational. When problems occur, it is usually due to reverse power relays and frequency-operated relays on grid lines due to severe grid disturbances. Under these circumstances, the emergency station supply must kick in to avoid damage to any equipment and to prevent hazardous situations such as the release of hydrogen gas from generators to the local environment.

In nuclear power plants

Emergency power systems, called there Emergency Diesel Generators (EDGs), are a required feature in nuclear power plants. They are typically installed in sets of three. The EDG installation is designed to the same safety-grade requirements as the other safety systems in the plant. The next (upcoming) generation of nuclear power plants includes some designs with multiple independent banks of EDGs (as in the ABWRs).

Controlling the emergency power system

For a 208 VAC emergency supply system, a central battery system with automatic controls, located in the power station building itself, is used to avoid long electric supply wires. This central battery system consists of lead-acid battery cell units to make up a 12 or 24 VDC system as well as stand-by cells, each with its own battery charging unit. Also needed are a voltage sensing unit capable of receiving 208 VAC and an automatic system that is able to signal to and activate the emergency supply circuit in case of failure of 208 VAC station supply.

Chapter-7

Fail-Safe

A **fail-safe** or **fail-secure** device is one that, in the event of failure, responds in a way that will cause no harm, or at least a minimum of harm, to other devices or danger to personnel.

Fail-safe components should not be confused with fail-secure components. A fail-secure component will allow, but does not cause, a system failure. For example, a fail-secure lock will remain locked during a failure and cannot be unlocked, even with the correct key. In contrast, a fail-safe component does not allow a system failure. For example, a lock that unlocks at the wrong time has failed, but is considered fail-safe because it does not open or attract undue attention to the door's unlocked state.

Significantly, despite popular belief to the contrary, "fail-safe" does not mean that failure is improbable, but rather that a system's design mitigates any unsafe consequences of failure.

Examples

Mechanical or physical

An aircraft lights its afterburners to maintain full power following an arrested landing aboard an aircraft carrier.

- Aircraft landing on an aircraft carrier increases the throttle to full power at touchdown. If the arresting wires fail to capture the plane, it is able to take off again.
- Coiling/rolling fire doors that are activated by building alarm systems or local smoke detectors must close automatically when signaled regardless of power. In case of power outage the coiling fire door does not need to close, but must be capable of automatic closing when given a signal from the building alarm systems or smoke detectors. A temperature sensitive fusible link may be employed to hold the fire doors open against gravity or a closing spring. In case of fire, the link melts and releases the doors, and they close.
- Luggage carts in airports in which the hand-brake must be held down at all times. If it is released, the cart will stop.
- Lawnmowers and snow blowers have a hand-closed lever that must be held down at all times. If it is released, it stops the blade's or rotor's rotation.
- Air brakes on railway trains and air brakes on trucks. The brakes are held in the 'off' position by air pressure created in the brake system. Should a brake line split, or a carriage become de-coupled, the air pressure will be lost and the brakes

applied. It is impossible to drive a train or truck with a serious leak in the air brake system.
- Motorized gates – In case of power outage the gate can be pushed open by hand with no crank or key required. However, as this would allow virtually anyone to go through the gate, a *fail-secure* design is used: In a power outage, the gate can only be opened by a hand crank that is usually kept in a safe area.
- During early Apollo program missions to the Moon, the spacecraft was put on a free return trajectory – if the engines failed at lunar orbit insertion, the craft would safely coast back to Earth.
- Elevator cabins that begin to accelerate too quickly as a result of the cables failing have a safety mechanism which uses contact with the guide rail to decelerate the car.
- Various devices that operate with fluids use fuses or valves as a fail-safe mechanism.
- A railway semaphore signal of the upper quadrant type is designed so that should the signal arm be weighed down by snow or the cable controlling the signal break the arm, returns to the 'danger' position, preventing any trains passing the inoperative signal.
- Diving watches – On diving watches the bezel is "unidirectional", i.e., it contains a ratchet so it can only be turned anti-clockwise to increase the apparent elapsed time. If the bezel could be turned the other way this could suggest to a diver that the elapsed time was shorter than the truth, thus giving a falsely low elapsed time reading and therefore an assumed falsely low air consumption reading and falsely high remaining air reading, all of which could be highly dangerous. In this fashion, the one-way bezel is designed to be "fail-safe", that is, if it 'fails', i.e., by being inadvertently rotated during the dive, it will only rotate so as to give a false reading of increased time below and thus less assumed tank air remaining rather than the opposite; it 'fails' in a 'safe' way.

Electrical or electronic

- Many devices are protected from short circuit with fuses. The destruction of the fuse will prevent destruction of the device.
- Avionics using redundant systems to perform the same computation with voting logic to determine the "safe" result.
- Traffic light controllers use a *Conflict Monitor Unit* to detect faults or conflicting signals and switch an intersection to all flashing red, rather than displaying potentially dangerous conflicting signals, e.g. showing green in all directions.
- The automatic protection of programs and/or processing systems when a hardware or software failure is detected in a computer system. A classic example is a watchdog timer.
- A control operation or function that prevents improper system functioning or catastrophic degradation in the event of circuit malfunction or operator error; for example, the **failsafe** track circuit used to control railway block signals.

- The iron pellet ballast on the Bathyscaphe is dropped to allow the submarine to ascend. The ballast is held in place by electromagnets. If electrical power fails the ballast is released, and the submarine then ascends to safety.
- Inside a modern CPU are features to prevent damage through overheating. In the event of cooling failure, the CPU will throttle then shut down beyond a critical temperature threshold to avoid damage.
- In industrial automation, alarm signals are usually "normally closed" (or active at 0). This insures that in case of a wire break the alarm will be triggered. If the signal were normally open, no wire failure would be detected.

Procedural

As well as physical devices and systems fail-safe procedures can be created so that if a procedure is not carried out or carried out incorrectly no dangerous action results. For example:

- In railway signalling signals which are not in active use for a train are required to be kept in the 'danger' position. The default position of every signal is therefore 'danger,' and therefore a positive action—setting signals to 'clear'—is required before a train may pass. This practice also ensures that, in case of a fault in the signalling system, an incapacitated signalman, or the unexpected entry of a train, that a train will never be shown an erroneous 'clear' signal.
- Train drivers are instructed that a railway signal showing a confusing, contradictory or unfamiliar aspect (for example a colour light signal that has suffered an electrical failure and is showing no light at all) must be treated as showing 'danger'. In this way, the driver contributes to the fail-safety of the system.

Other terminology

Fail-safe (foolproof) devices are also known as *poka-yoke* devices. *Poka-yoke*, a Japanese term, was coined by Shigeo Shingo, a quality guru.

Chapter-8

Fly-by-Wire

Fly-by-wire

Green colored flight control wiring of a test aircraft

A **fly-by-wire** (FBW) system replaces manual flight control of an aircraft with an electronic interface. The movements of flight controls are converted to electronic signals transmitted by wires (hence the fly-by-wire term), and flight control computers determine how to move the actuators at each control surface to provide the ordered response. The fly-by-wire system also allows automatic signals sent by the aircraft's computers to perform functions without the pilot's input, as in systems that automatically help stabilize the aircraft.

Development

Mechanical and hydro-mechanical flight control systems are relatively heavy and require careful routing of flight control cables through the aircraft by systems of pulleys, cranks, tension cables and hydraulic pipes. Both systems often require redundant backup to deal with failures, which again increases weight. Furthermore, both have limited ability to compensate for changing aerodynamic conditions. Dangerous characteristics such as stalling, spinning and pilot-induced oscillation (PIO), which depend mainly on the

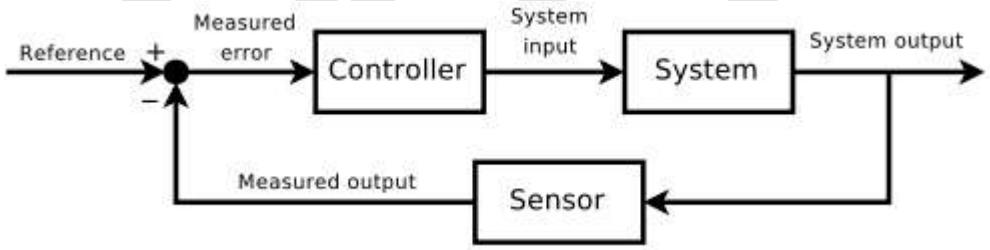
stability and structure of the aircraft concerned rather than the control system itself, can still occur with these systems.

The term "fly-by-wire" implies a purely electrically-signaled control system. However, it is used in the general sense of computer-configured controls, where a computer system is interposed between the operator and the final control actuators or surfaces. This modifies the manual inputs of the pilot in accordance with control parameters.

Side-sticks, center sticks, or conventional flight control yokes can be used to fly FBW aircraft. While the side-stick offers the advantages of being lighter, mechanically simpler, and unobtrusive, The Boeing Company's aerospace engineers decided that the lack of visual feedback (none given by side-sticks) is a significant problem, and so they designed conventional control yokes in the Boeing 777 and the brand-new Boeing 787, which is undergoing flight tests as of June 2010. This same approach has been used for the Embraer 170/190 jets. Most Airbus airliners are operated with side-sticks.

Basic operation

Command



Simple feedback loop

Fly-by wire systems are by their nature quite complex however their operation can be explained in relatively simple terms. When a pilot moves the control column (or sidestick) a signal is sent to a computer, this is analogous to moving a game controller, the signal is sent through multiple wires (channels) to ensure that the signal reaches the computer. When there are three channels being used this is known as 'Triplex'. The computer receives the signals, performs a calculation (adds the signal voltages and divides by the number of signals received to find the mean average voltage) and adds another channel. These four 'Quadruplex' signals are then sent to the control surface actuator and the surface begins to move. Potentiometers in the actuator send a signal back to the computer (usually a negative voltage) reporting the position of the actuator. When the actuator reaches the desired position the two signals (incoming and outgoing) cancel each other out and the actuator stops moving (completing a feedback loop).

Automatic Stability Systems

Fly-by-wire control systems allow aircraft computers to perform tasks without pilot input. Automatic stability systems operate in this way. Gyroscopes fitted with sensors are

mounted in an aircraft to sense movement changes in the pitch, roll and yaw axes. Any movement (from straight and level flight for example) results in signals to the computer, which automatically moves control actuators to stabilize the aircraft.

Safety and redundancy

Aircraft systems may be quadruplexed (four independent channels) to prevent loss of signals in the case of failure of one or even two channels. High performance aircraft that have FBW controls (also called CCVs or Control-Configured Vehicles) may be deliberately designed to have low or even negative aerodynamic stability in some flight regimes, the rapid-reacting CCV controls compensating for the lack of natural stability.

Pre-flight safety checks of a fly-by-wire system are often performed using Built-In Test Equipment (BITE). On programming the system, either by the pilot or groundcrew, a number of control movement steps are automatically performed. Any failure will be indicated to the crews.

Some aircraft, the Panavia Tornado for example, retain a very basic hydro-mechanical backup system for limited flight control capability on losing electrical power, in the case of the Tornado this allows rudimentary control of the stabilators only for pitch and roll axis movements.

Weight Saving

A FBW aircraft can be lighter than a similar design with conventional controls. Partly due to the lower overall weight of the system components; and partly because the natural aerodynamic stability of the aircraft can be relaxed, slightly for a transport aircraft and more for a maneuverable fighter, which means that the stability surfaces that are part of the aircraft structure can therefore be made smaller. These include the vertical and horizontal stabilizers (fin and tailplane) that are (normally) at the rear of the fuselage. If these structures can be reduced in size, airframe weight is reduced. The advantages of FBW controls were first exploited by the military and then in the commercial airline market. The Airbus series of airliners used full-authority FBW controls beginning with their A320 series. Boeing followed with their 777 and later designs.

Electronic fly-by-wire systems can respond flexibly to changing aerodynamic conditions, by tailoring flight control surface movements so that aircraft response to control inputs is appropriate to flight conditions. Electronic systems require less maintenance, whereas mechanical and hydraulic systems require lubrication, tension adjustments, leak checks, fluid changes, etc. Furthermore, putting circuitry between pilot and aircraft can enhance safety; for example the control system can try to prevent a stall, or it can stop the pilot from over stressing the airframe.

The main concern with fly-by-wire systems is reliability. While traditional mechanical or hydraulic control systems usually fail gradually, the loss of all flight control computers could immediately render the aircraft uncontrollable. For this reason, most fly-by-wire

systems incorporate either redundant computers (triplex, quadruplex etc.), some kind of mechanical or hydraulic backup or a combination of both. A "mixed" control system such as the latter is not desirable and modern FBW aircraft normally avoid it by having more independent FBW channels, thereby reducing the possibility of overall failure to minuscule levels that are acceptable to the independent regulatory and safety authority responsible for aircraft design, testing and certification before operational service.

History

F-8C Crusader digital fly-by-wire testbed

Electronic signalling of the control surfaces was tested in the 1950s. This replaced long runs of mechanical and hydraulic connections with electrical ones.

The first non-experimental aircraft that was designed and flown (in 1958) with a fly-by-wire flight control system was the Avro Canada CF-105 Arrow. a feat not repeated with a production aircraft until Concorde in 1969. This system also included solid-state components and system redundancy, was designed to be integrated with a computerised navigation and automatic search and track radar, was flyable from ground control with data uplink and downlink, and provided artificial feel (feedback) to the pilot.

The first digital fly-by-wire aircraft to take to the air (in 1972) was an F-8 Crusader, which had been modified electronically by the National Aeronautics and Space Administration of the United States as a test aircraft, a feat mirrored in the USSR by the Sukhoi T-4. At about the same time in the United Kingdom a trainer variant of the British Hawker Hunter fighter was modified at the British Royal Aircraft Establishment with fly-by-wire flight controls for the right-seat pilot. This was test-flown, with the left-seat pilot

having conventional flight controls for safety reasons, and with the capability for him to override and turn off the fly-by-wire system.

Analog systems

All "fly-by-wire" flight control systems eliminate the complexity, the fragility, and the weight of the mechanical circuit of the hydromechanical or electromechanical flight control systems. Fly-by-wire replace those with electronic circuits. The control mechanisms in the cockpit now operate signal transducers, which in turn generate the appropriate electronic commands. These are next processed by an electronic controller, either an analog one, or more modernly, a digital one. Aircraft and spacecraft autopilots are now part of the electronic controller.

The hydraulic circuits are similar except that mechanical servo valves are replaced with electrically-controlled servo valves, operated by the electronic controller. This is the simplest and earliest configuration of an analog fly-by-wire flight control system. In this configuration, the flight control systems must simulate "feel". The electronic controller controls electrical feel devices that provide the appropriate "feel" forces on the manual controls. This was used in Concorde, the first production fly-by-wire airliner.

In more sophisticated versions, analog computers replaced the electronic controller. The canceled 1950s Canadian supersonic intercepter, the Avro Canada CF-105 Arrow, employed this type of system. Analog computers also allowed some customization of flight control characteristics, including relaxed stability. This was exploited by the early versions of F-16, giving it impressive maneuverability.

Digital systems

The Airbus A320, first airliner with digital fly-by-wire controls

A digital fly-by-wire flight control system is similar to its analog counterpart. However, the signal processing is done by digital computers and the pilot literally can "fly-via-computer". This also increases the flexibility of the flight control system, since the digital computers can receive input from any aircraft sensor (such as the altimeters and the pitot tubes. This also increases the electronic stability, because the system is less dependent on the values of critical electrical components in an analog controller.

The computers sense position and force inputs from pilot controls and aircraft sensors. They solve differential equations to determine the appropriate command signals that move the flight controls to execute the intentions of the pilot.

The programming of the digital computers enable flight envelope protection. In this aircraft designers precisely tailor an aircraft's handling characteristics, to stay within the overall limits of what is possible given the aerodynamics and structure of the aircraft. For example, the computer in flight envelope protection mode can try to prevent the aircraft from being handled dangerously by preventing pilots from exceeding preset limits on the aircraft's flight-control envelope, such as those that prevent stalls and spins, and which limit airspeeds and g forces on the airplane. Software can also be included that stabilize the flight-control inputs to avoid pilot-induced oscillations.

Since the flight-control computers continuously "fly" the aircraft, pilot's workloads can be reduced. Also, in military and naval applications, it is now possible to fly military aircraft that have relaxed stability. The primary benefit for such aircraft is more maneuverability during combat and training flights, and the so-called "carefree handling" because stalling, spinning. and other undesirable performances are prevented automatically by the computers.

Digital flight control systems enable inherently unstable combat aircraft, such as the F-117 Nighthawk and the B-2 Spirit flying wing to fly in usable and safe manners.

Applications

A Dassault Falcon 7X, the first business jet with digital fly-by-wire controls

- The Space Shuttle Orbiter has an all-digital fly-by-wire control system. This system was first exercised (as the only flight control system) during the glider unpowered-flight "Approach and Landing Tests" that began on the Space Shuttle *Enterprise* during 1977.
- During 1984, the Airbus Industries Airbus A320 became the first airliner to fly with an all-digital fly-by-wire control system.
- During 2005, the Dassault Falcon 7X became the first business jet with fly-by-wire controls.

Legislation

The Federal Aviation Administration (FAA) of the United States has adopted the RTCA/DO-178B, titled "Software Considerations in Airborne Systems and Equipment Certification", as the certification standard for aviation software. Any safety-critical component in a digital fly-by-wire system including applications of the laws of aeronautics and computer operating systems will need to be certified to DO-178B Level A, which is applicable for preventing potential catastrophic failures.

Nevertheless, the top concern for computerized, digital, fly-by-wire systems is reliability, even more so than for analog electronic control systems. This is because the digital computers that are running software are often the only control path between the pilot and

aircraft's flight control surfaces. If the computer software crashes for any reason, the pilot may be unable to control an aircraft. Hence virtually all fly-by-wire flight control systems are either triply or quadruply redundant in their computers and electronics. These have three or four flight-control computers operating in parallel, and three or four separate data buses connecting them with each control surface.

Redundancy

If one of the flight-control computers crashes, or is damaged in combat, or suffers from "insanity" caused by electromagnetic pulses, the others overrule the faulty one (or even two of them), they continue flying the aircraft safely, and they can either turn off or re-boot the faulty computers. Any flight-control computer whose results disagree with the others is ruled to be faulty, and it is either ignored or re-booted. (In other words, it is voted-out of control by the others.)

In addition, most of the early digital fly-by-wire aircraft also had an analog electrical, a mechanical, or a hydraulic back-up flight control system. The Space Shuttle has, in addition to its redundant set of four digital computers running its primary flight-control software, a fifth back-up computer running a separately developed, reduced-function, software flight-control system - one that can be commanded to take over in the event that a fault ever affects all of the computers in the other four. This back-up system serves to reduce the risk of total flight-control-system failure ever happening because of a general-purpose flight software fault has escaped notice in the other four computers.

For airliners, flight-control redundancy improves their safety, but fly-by-wire control systems also improve economy in flight because they are lighter, and they eliminate the need for many mechanical, and heavy, flight-control mechanisms. Furthermore, most modern airliners have computerized systems that control their jet engine throttles, air inlets, fuel storage and distribution system, in such a way to minimize their consumption of jet fuel. Thus, digital control systems do their best to reduce the cost of flights.

Airbus/Boeing

Airbus and Boeing commercial airplanes differ in their approaches in using fly-by-wire systems. In Airbus airliners, the flight-envelope control system always retains ultimate flight control, and it will not permit the pilots to fly outside these performance limits. However, in the event of multiple failures of redundant computers, the A320 does have mechanical back-up system for its pitch trim and its rudder. The A340-600 has a purely electrical (not electronic) back-up rudder control system, and beginning with the new A380 airliner, all flight-control systems have back-up systems that are purely electrical through the use of a so-called "three-axis Backup Control Module" (BCM)

With the Boeing 777 model airliners, the two pilots can completely override the computerized flight-control system to permit the aircraft to be flown beyond its usual flight-control envelope during emergencies. Airbus's strategy, which began with the Airbus A320, has been continued on subsequent Airbus airliners.

Engine digital control

The advent of FADEC (Full Authority Digital Engine Control) engines permits operation of the flight control systems and autothrottles for the engines to be fully integrated. On modern military aircraft other systems such as autostabilization, navigation, radar and weapons system are all integrated with the flight control systems. FADEC allows maximum performance to be extracted from the aircraft without fear of engine misoperation, aircraft damage or high pilot workloads. In the civil field, the integration increases flight safety and economy. The Airbus A320 and its fly-by-wire brethren are protected from dangerous situations such as low-speed stall or overstressing by flight envelope protection. As a result, in such conditions, the flight control systems commands the engines to increase thrust without pilot intervention. In economy cruise modes, the flight control systems adjust the throttles and fuel tank selections more precisely than all but the most skillful pilots. FADEC reduces rudder drag needed to compensate for sideways flight from unbalanced engine thrust. On the A330/A340 family, fuel is transferred between the main (wing and center fuselage) tanks and a fuel tank in the horizontal stabilizer, to optimize the aircraft's center of gravity during cruise flight. The fuel management controls keep the aircraft's center of gravity accurately trimmed with fuel weight, rather than drag-inducing aerodynamic trims in the elevators.

Further developments

Fly-by-optics

Fly-by-optics is sometimes used instead of fly-by-wire because it can transfer data at higher speeds, and it is immune to electromagnetic interference. In most cases, the cables are just changed from electrical to optical fiber cables. Sometimes it is referred to as "fly-by-light" due to its use of fiber optics. The data generated by the software and interpreted by the controller remain the same.

Power-by-wire

Having eliminated the mechanical transmission circuits in fly-by-wire flight control systems, the next step is to eliminate the bulky and heavy hydraulic circuits. The hydraulic circuit is replaced by an electrical power circuit. The power circuits power electrical or self-contained electrohydraulic actuators that are controlled by the digital flight control computers. All benefits of digital fly-by-wire are retained.

The biggest benefits are weight savings, the possibility of redundant power circuits and tighter integration between the aircraft flight control systems and its avionics systems. The absence of hydraulics greatly reduces maintenance costs. This system is used in the Lockheed Martin F-35 Lightning II and in Airbus A380 backup flight controls. The Boeing 787 will also incorporate some electrically operated flight controls (spoilers and horizontal stabilizer), which will remain operational with either a total hydraulics failure and/or flight control computer failure.

Fly-by-wireless

Wiring adds a considerable amount of weight to an aircraft; therefore, researchers are exploring implementing fly-by-wireless solutions. Fly-by-wireless systems are very similar to fly-by-wire systems, however, instead of using a wired protocol for the physical layer a wireless protocol is employed.

In addition to reducing weight, implementing a wireless solution has the potential to reduce costs throughout an aircraft's life cycle. For example, many key failure points associated with wire and connectors will be eliminated thus hours spent troubleshooting wires and connectors will be reduced. Furthermore, engineering costs could potentially decrease because less time would be spent on designing wiring installations, late changes in an aircraft's design would be easier to manage, etc.

Intelligent Flight Control System

A newer flight control system, called Intelligent Flight Control System (IFCS), is an extension of modern digital fly-by-wire flight control systems. The aim is to intelligently compensate for aircraft damage and failure during flight, such as automatically using engine thrust and other avionics to compensate for severe failures such as loss of hydraulics, loss of rudder, loss of ailerons, loss of an engine, etc. Several demonstrations were made on a flight simulator where a Cessna-trained small-aircraft pilot successfully landed a heavily-damaged full-size concept jet, without prior experience with large-body jet aircraft. This development is being spearheaded by NASA Dryden Flight Research Center. It is reported that enhancements are mostly software upgrades to existing fully computerized digital fly-by-wire flight control systems.

Chapter-9

Error Detection and Correction

In information theory and coding theory with applications in computer science and telecommunication, **error detection and correction** or **error control** are techniques that enable reliable delivery of digital data over unreliable communication channels. Many communication channels are subject to channel noise, and thus errors may be introduced during transmission from the source to a receiver. Error detection techniques allow detecting such errors, while error correction enables reconstruction of the original data.

The general definitions of the terms are as follows:

- *Error detection* is the detection of errors caused by noise or other impairments during transmission from the transmitter to the receiver.
- *Error correction* is the detection of errors and reconstruction of the original, error-free data.

Error correction may generally be realized in two different ways:

- *Automatic repeat request (ARQ)* (sometimes also referred to as *backward error correction*): This is an error control technique whereby an error detection scheme is combined with requests for retransmission of erroneous data. Every block of data received is checked using the error detection code used, and if the check fails, retransmission of the data is requested – this may be done repeatedly, until the data can be verified.
- *Forward error correction (FEC)*: The sender encodes the data using an *error-correcting code (ECC)* prior to transmission. The additional information (redundancy) added by the code is used by the receiver to recover the original data. In general, the reconstructed data is what is deemed the "most likely" original data.

ARQ and FEC may be combined, such that minor errors are corrected without retransmission, and major errors are corrected via a request for retransmission: this is called *hybrid automatic repeat-request (HARQ)*.

Introduction

The general idea for achieving error detection and correction is to add some redundancy (i.e., some extra data) to a message, which receivers can use to check consistency of the delivered message, and to recover data determined to be erroneous. Error-detection and correction schemes can be either systematic or non-systematic: In a systematic scheme, the transmitter sends the original data, and attaches a fixed number of *check bits* (or *parity data*), which are derived from the data bits by some deterministic algorithm. If only error detection is required, a receiver can simply apply the same algorithm to the received data bits and compare its output with the received check bits; if the values do not match, an error has occurred at some point during the transmission. In a system that uses a non-systematic code, the original message is transformed into an encoded message that has at least as many bits as the original message.

Good error control performance requires the scheme to be selected based on the characteristics of the communication channel. Common channel models include memory-less models where errors occur randomly and with a certain probability, and dynamic models where errors occur primarily in bursts. Consequently, error-detecting and correcting codes can be generally distinguished between *random-error-detecting/correcting* and *burst-error-detecting/correcting*. Some codes can also be suitable for a mixture of random errors and burst errors.

If the channel capacity cannot be determined, or is highly varying, an error-detection scheme may be combined with a system for retransmissions of erroneous data. This is known as automatic repeat request (ARQ), and is most notably used in the Internet. An alternate approach for error control is hybrid automatic repeat request (HARQ), which is a combination of ARQ and error-correction coding.

Error detection schemes

Error detection is most commonly realized using a suitable hash function (or checksum algorithm). A hash function adds a fixed-length *tag* to a message, which enables receivers to verify the delivered message by recomputing the tag and comparing it with the one provided.

There exists a vast variety of different hash function designs. However, some are of particularly widespread use because of either their simplicity or their suitability for detecting certain kinds of errors (e.g., the cyclic redundancy check's performance in detecting burst errors).

Random-error-correcting codes based on minimum distance coding can provide a suitable alternative to hash functions when a strict guarantee on the minimum number of errors to be detected is desired. Repetition codes, described below, are special cases of error-correcting codes: although rather inefficient, they find applications for both error correction and detection due to their simplicity.

Repetition codes

A **repetition code** is a coding scheme that repeats the bits across a channel to achieve error-free communication. Given a stream of data to be transmitted, the data is divided into blocks of bits. Each block is transmitted some predetermined number of times. For example, to send the bit pattern "1011", the four-bit block can be repeated three times, thus producing "1011 1011 1011". However, if this twelve-bit pattern was received as "1010 1011 1011" – where the first block is unlike the other two – it can be determined that an error has occurred.

Repetition codes are not very efficient, and can be susceptible to problems if the error occurs in exactly the same place for each group (e.g., "1010 1010 1010" in the previous example would be detected as correct). The advantage of repetition codes is that they are extremely simple, and are in fact used in some transmissions of numbers stations.

Parity bits

A **parity bit** is a bit that is added to a group of source bits to ensure that the number of set bits (i.e., bits with value 1) in the outcome is even or odd. It is a very simple scheme that can be used to detect single or any other odd number (i.e., three, five, etc.) of errors in the output. An even number of flipped bits will make the parity bit appear correct even though the data is erroneous.

Extensions and variations on the parity bit mechanism are horizontal redundancy checks, vertical redundancy checks, and "double," "dual," or "diagonal" parity (used in RAID-DP).

Checksums

A **checksum** of a message is a modular arithmetic sum of message code words of a fixed word length (e.g., byte values). The sum may be negated by means of a one's-complement prior to transmission to detect errors resulting in all-zero messages.

Checksum schemes include parity bits, check digits, and longitudinal redundancy checks. Some checksum schemes, such as the Luhn algorithm and the Verhoeff algorithm, are specifically designed to detect errors commonly introduced by humans in writing down or remembering identification numbers.

Cyclic redundancy checks (CRCs)

A **cyclic redundancy check (CRC)** is a single-burst-error-detecting cyclic code and non-secure hash function designed to detect accidental changes to digital data in computer networks. It is characterized by specification of a so-called *generator polynomial*, which is used as the divisor in a polynomial long division over a finite field, taking the input data as the dividend, and where the remainder becomes the result.

Cyclic codes have favorable properties in that they are well suited for detecting burst errors. CRCs are particularly easy to implement in hardware, and are therefore commonly used in digital networks and storage devices such as hard disk drives.

Even parity is a special case of a cyclic redundancy check, where the single-bit CRC is generated by the divisor $x+1$.

Cryptographic hash functions

A **cryptographic hash function** can provide strong assurances about data integrity, provided that changes of the data are only accidental (i.e., due to transmission errors). Any modification to the data will likely be detected through a mismatching hash value. Furthermore, given some hash value, it is infeasible to find some input data (other than the one given) that will yield the same hash value. Message authentication codes, also called *keyed* cryptographic hash functions, provide additional protection against intentional modification by an attacker.

Error-correcting codes

Any error-correcting code can be used for error detection. A code with *minimum Hamming distance*, d, can detect up to $d-1$ errors in a code word. Using minimum-distance-based error-correcting codes for error detection can be suitable if a strict limit on the minimum number of errors to be detected is desired.

Codes with minimum Hamming distance $d=2$ are degenerate cases of error-correcting codes, and can be used to detect single errors. The parity bit is an example of a single-error-detecting code.

The Berger code is an early example of a unidirectional error(-correcting) code that can detect any number of errors on an asymmetric channel, provided that only transitions of cleared bits to set bits *or* set bits to cleared bits can occur.

Error correction

Automatic repeat request

Automatic Repeat reQuest (ARQ) is an error control method for data transmission that makes use of error-detection codes, acknowledgment and/or negative acknowledgment messages, and timeouts to achieve reliable data transmission. An *acknowledgment* is a message sent by the receiver to indicate that it has correctly received a data frame.

Usually, when the transmitter does not receive the acknowledgment before the timeout occurs (i.e., within a reasonable amount of time after sending the data frame), it retransmits the frame until it is either correctly received or the error persists beyond a predetermined number of retransmissions.

Three types of ARQ protocols are Stop-and-wait ARQ, Go-Back-N ARQ, and Selective Repeat ARQ.

ARQ is appropriate if the communication channel has varying or unknown capacity, such as is the case on the Internet. However, ARQ requires the availability of a back channel, results in possibly increased latency due to retransmissions, and requires the maintenance of buffers and timers for retransmissions, which in the case of network congestion can put a strain on the server and overall network capacity.

Error-correcting code

An error-correcting code (ECC) or forward error correction (FEC) code is a system of adding redundant data, or *parity data*, to a message, such that it can be recovered by a receiver even when a number of errors (up to the capability of the code being used) were introduced, either during the process of transmission, or on storage. Since the receiver does not have to ask the sender for retransmission of the data, a back-channel is not required in forward error correction, and it is therefore suitable for simplex communication such as broadcasting. Error-correcting codes are frequently used in lower-layer communication, as well as for reliable storage in media such as CDs, DVDs, hard disks, and RAM.

Error-correcting codes are usually distinguished between convolutional codes and block codes:

- *Convolutional codes* are processed on a bit-by-bit basis. They are particularly suitable for implementation in hardware, and the Viterbi decoder allows optimal decoding.

- *Block codes* are processed on a block-by-block basis. Early examples of block codes are repetition codes, Hamming codes and multidimensional parity-check codes. They were followed by a number of efficient codes, Reed-Solomon codes being the most notable due to their current widespread use. Turbo codes and low-density parity-check codes (LDPC) are relatively new constructions that can provide almost optimal efficiency.

Shannon's theorem is an important theorem in forward error correction, and describes the maximum information rate at which reliable communication is possible over a channel that has a certain error probability or signal-to-noise ratio (SNR). This strict upper limit is expressed in terms of the channel capacity. More specifically, the theorem says that there exist codes such that with increasing encoding length the probability of error on a discrete memoryless channel can be made arbitrarily small, provided that the code rate is smaller than the channel capacity. The code rate is defined as the fraction k/n of k source symbols and n encoded symbols.

The actual maximum code rate allowed depends on the error-correcting code used, and may be lower. This is because Shannon's proof was only of existential nature, and did not

show how to construct codes which are both optimal and have efficient encoding and decoding algorithms.

Hybrid schemes

Hybrid ARQ is a combination of ARQ and forward error correction. There are two basic approaches:

- Messages are always transmitted with FEC parity data (and error-detection redundancy). A receiver decodes a message using the parity information, and requests retransmission using ARQ only if the parity data was not sufficient for successful decoding (identified through a failed integrity check).
- Messages are transmitted without parity data (only with error-detection information). If a receiver detects an error, it requests FEC information from the transmitter using ARQ, and uses it to reconstruct the original message.

The latter approach is particularly attractive on an erasure channel when using a rateless erasure code.

Applications

Applications that require low latency (such as telephone conversations) cannot use Automatic Repeat reQuest (ARQ); they must use Forward Error Correction (FEC). By the time an ARQ system discovers an error and re-transmits it, the re-sent data will arrive too late to be any good.

Applications where the transmitter immediately forgets the information as soon as it is sent (such as most television cameras) cannot use ARQ; they must use FEC because when an error occurs, the original data is no longer available. (This is also why FEC is used in data storage systems such as RAID and distributed data store).

Applications that use ARQ must have a return channel. Applications that have no return channel cannot use ARQ.

Applications that require extremely low error rates (such as digital money transfers) must use ARQ.

The Internet

In a typical TCP/IP stack, error control is performed at multiple levels:

- Each Ethernet frame carries a CRC-32 checksum. Frames received with incorrect checksums are discarded by the receiver hardware.
- The IPv4 header contains a checksum protecting the contents of the header. Packets with mismatching checksums are dropped within the network or at the receiver.

- The checksum was omitted from the IPv6 header in order to minimize processing costs in network routing and because current link layer technology is assumed to provide sufficient error detection.
- UDP has an optional checksum covering the payload and addressing information from the UDP and IP headers. Packets with incorrect checksums are discarded by the operating system network stack. The checksum is optional under IPv4, only, because the IP layer checksum may already provide the desired level of error protection.
- TCP provides a checksum for protecting the payload and addressing information from the TCP and IP headers. Packets with incorrect checksums are discarded within the network stack, and eventually get retransmitted using ARQ, either explicitly (such as through triple-ack) or implicitly due to a timeout.

Deep-space telecommunications

Development of error-correction codes was tightly coupled with the history of deep-space missions due to the extreme dilution of signal power over interplanetary distances, and the limited power availability aboard space probes. Whereas early missions sent their data uncoded, starting from 1968 digital error correction was implemented in the form of (sub-optimally decoded) convolutional codes and Reed-Muller codes. The Reed-Muller code was well suited to the noise the spacecraft was subject to (approximately matching a bell curve), and was implemented at the Mariner spacecraft for missions between 1969 and 1977.

The Voyager 1 and Voyager 2 missions, which started in 1977, were designed to deliver color imaging amongst scientific information of Jupiter and Saturn. This resulted in increased coding requirements, and thus the spacecraft were supported by (optimally Viterbi-decoded) convolutional codes that could be concatenated with an outer Golay (24,12,8) code. The Voyager 2 probe additionally supported an implementation of a Reed-Solomon code: the concatenated Reed-Solomon-Viterbi (RSV) code allowed for very powerful error correction, and enabled the spacecraft's extended journey to Uranus and Neptune.

The CCSDS currently recommends usage of error correction codes with performance similar to the Voyager 2 RSV code as a minimum. Concatenated codes are increasingly falling out of favor with space missions, and are replaced by more powerful codes such as Turbo codes or LDPC codes.

The different kinds of deep space and orbital missions that are conducted suggest that trying to find a "one size fits all" error correction system will be an ongoing problem for some time to come. For missions close to earth the nature of the channel noise is different from that of a spacecraft on an interplanetary mission experiences. Additionally, as a spacecraft increases its distance from earth, the problem of correcting for noise gets larger.

Satellite broadcasting (DVB)

The demand for satellite transponder bandwidth continues to grow, fueled by the desire to deliver television (including new channels and High Definition TV) and IP data. Transponder availability and bandwidth constraints have limited this growth, because transponder capacity is determined by the selected modulation scheme and Forward error correction (FEC) rate.

Overview

- QPSK coupled with traditional Reed Solomon and Viterbi codes have been used for nearly 20 years for the delivery of digital satellite TV.
- Higher order modulation schemes such as 8PSK, 16QAM and 32QAM have enabled the satellite industry to increase transponder efficiency by several orders of magnitude.
- This increase in the information rate in a transponder comes at the expense of an increase in the carrier power to meet the threshold requirement for existing antennas.
- Tests conducted using the latest chipsets demonstrate that the performance achieved by using Turbo Codes may be even lower than the 0.8 dB figure assumed in early designs.

Data storage

Error detection and correction codes are often used to improve the reliability of data storage media.

A "parity track" was present on the first magnetic tape data storage in 1951. The "Optimal Rectangular Code" used in group code recording tapes not only detects but also corrects single-bit errors.

Some file formats, particularly archive formats, include a checksum (most often CRC32) to detect corruption and truncation and can employ redundancy and/or parity files to recover portions of corrupted data.

Reed Solomon codes are used in compact discs to correct errors caused by scratches.

Modern hard drives use CRC codes to detect and Reed-Solomon codes to correct minor errors in sector reads, and to recover data from sectors that have "gone bad" and store that data in the spare sectors.

RAID systems use a variety of error correction techniques, to correct errors when a hard drive completely fails.

Error-correcting memory

DRAM memory may provide increased protection against soft errors by relying on error correcting codes. Such error-correcting memory, known as *ECC* or *EDAC-protected* memory, is particularly desirable for high fault-tolerant applications, such as servers, as well as deep-space applications due to increased radiation.

Error-correcting memory controllers traditionally use Hamming codes, although some use triple modular redundancy.

Interleaving allows distributing the effect of a single cosmic ray potentially upsetting multiple physically neighboring bits across multiple words by associating neighboring bits to different words. As long as a single event upset (SEU) does not exceed the error threshold (e.g., a single error) in any particular word between accesses, it can be corrected (e.g., by a single-bit error correcting code), and the illusion of an error-free memory system may be maintained.

Chapter-10

Coding Theory

Coding theory is the study of the properties of codes and their fitness for a specific application. Codes are used for data compression, cryptography, error-correction and more recently also for network coding. Codes are studied by various scientific disciplines—such as information theory, electrical engineering, mathematics, and computer science—for the purpose of designing efficient and reliable data transmission methods. This typically involves the removal of redundancy and the correction (or detection) of errors in the transmitted data.

There are essentially two aspects to Coding theory:

1. Data compression (or, *source coding*)
2. Error correction (or, *channel coding*).

These two aspects may be studied in combination. Source encoding, attempts to compress the data from a source in order to transmit it more efficiently. This practice is found every day on the Internet where the common "Zip" data compression is used to reduce the network load and make files smaller. The second, channel encoding, adds extra data bits to make the transmission of data more robust to disturbances present on the transmission channel. The ordinary user may not be aware of many applications using channel coding. A typical music CD uses the Reed-Solomon code to correct for scratches and dust. In this application the transmission channel is the CD itself. Cell phones also use coding techniques to correct for the fading and noise of high frequency radio transmission. Data modems, telephone transmissions, and NASA all employ channel coding techniques to get the bits through, for example the turbo code and LDPC codes.

Source coding

The aim of source coding is to take the source data and make it smaller.

Principle

Entropy of a source is the measure of information. Basically source codes try to reduce the redundancy present in the source, and represent the source with fewer bits that carry more information.

Data compression which explicitly tries to minimize the average length of messages according to a particular assumed probability model is called entropy encoding.

Various techniques used by source coding schemes try to achieve the limit of Entropy of the source. $C(x) \geq H(x)$, where $H(x)$ is entropy of source (bitrate), and $C(x)$ is the bitrate after compression. In particular, no source coding scheme can be better than the entropy of the source.

Example

Facsimile transmission uses a simple run length code. Source coding includes also removal of all data that superfluous the need of transmitter, this decreases the bandwidth required for the transmission process.

Channel coding

The aim of channel coding theory is to find codes which transmit quickly, contain many valid code words and can correct or at least detect many errors. While not mutually exclusive, performance in these areas is a trade off. So, different codes are optimal for different applications. The needed properties of this code mainly depend on the probability of errors happening during transmission. In a typical CD, the impairment is mainly dust or scratches. Thus codes are used in an interleaved manner. The data is spread out over the disk. Although not a very good code, a simple repeat code can serve as an understandable example. Suppose we take a block of data bits (representing sound) and send it three times. At the receiver we will examine the three repetitions bit by bit and take a majority vote. The twist on this is that we don't merely send the bits in order. We interleave them. The block of data bits is first divided into 4 smaller blocks. Then we cycle through the block and send one bit from the first, then the second, etc. This is done three times to spread the data out over the surface of the disk. In the context of the simple repeat code, this may not appear effective. However, there are more powerful codes known which are very effective at correcting the "burst" error of a scratch or a dust spot when this interleaving technique is used.

Other codes are more appropriate for different applications. Deep space communications are limited by the thermal noise of the receiver which is more of a continuous nature than a bursty nature. Likewise, narrowband modems are limited by the noise, present in the telephone network and also modeled better as a continuous disturbance. Cell phones are subject to rapid fading. The high frequencies used can cause rapid fading of the signal even if the receiver is moved a few inches. Again there are a class of channel codes that are designed to combat fading.

Linear codes

The term **algebraic coding theory** denotes the sub-field of coding theory where the properties of codes are expressed in algebraic terms and then further researched.

Algebraic coding theory is basically divided into two major types of codes:

1. Linear block codes
2. Convolutional codes.

It analyzes the following three properties of a code – mainly:

- code word length
- total number of valid code words
- the minimum distance between two valid code words, using mainly the Hamming distance, sometimes also other distances like the Lee distance.

Linear block codes

Linear block codes have the property of linearity, i.e the sum of any two codewords is also a code word, and they are applied to the source bits in blocks, hence the name linear block codes. There are block codes that are not linear, but it is difficult to prove that a code is a good one without this property.

Linear block codes are summarized by their symbol alphabets (e.g., binary or ternary) and parameters (n,m,d_{min}) where

1. n is the length of the codeword, in symbols,
2. m is the number of source symbols that will be used for encoding at once,
3. d_{min} is the minimum hamming distance for the code.

There are many types of linear block codes, such as

1. Cyclic codes (e.g., Hamming codes)
2. Repetition codes
3. Parity codes
4. Polynomial codes (e.g., BCH codes)
5. Reed–Solomon codes
6. Algebraic geometric codes
7. Reed–Muller codes
8. Perfect codes.

Block codes are tied to the sphere packing problem, which has received some attention over the years. In two dimensions, it is easy to visualize. Take a bunch of pennies flat on the table and push them together. The result is a hexagon pattern like a bee's nest. But block codes rely on more dimensions which cannot easily be visualized. The powerful

(24,12) Golay code used in deep space communications uses 24 dimensions. If used as a binary code (which it usually is) the dimensions refer to the length of the codeword as defined above.

The theory of coding uses the *N*-dimensional sphere model. For example, how many pennies can be packed into a circle on a tabletop, or in 3 dimensions, how many marbles can be packed into a globe. Other considerations enter the choice of a code. For example, hexagon packing into the constraint of a rectangular box will leave empty space at the corners. As the dimensions get larger, the percentage of empty space grows smaller. But at certain dimensions, the packing uses all the space and these codes are the so-called "perfect" codes. The only nontrivial and useful perfect codes are the distance-3 Hamming codes with parameters satisfying $(2^r - 1, 2^r - 1 - r, 3)$, and the [23,12,7] binary and [11,6,5] ternary Golay codes.

Another code property is the number of neighbors that a single codeword may have. Again, consider pennies as an example. First we pack the pennies in a rectangular grid. Each penny will have 4 near neighbors (and 4 at the corners which are farther away). In a hexagon, each penny will have 6 near neighbors. When we increase the dimensions, the number of near neighbors increases very rapidly. The result is the number of ways for noise to make the receiver choose a neighbor (hence an error) grows as well. This is a fundamental limitation of block codes, and indeed all codes. It may be harder to cause an error to a single neighbor, but the number of neighbors can be large enough so the total error probability actually suffers.

Properties of linear block codes are used in many applications. For example, the syndrome-coset uniqueness property of linear block codes is used in trellis shaping, one of the best known shaping codes. This same property is used in sensor networks for distributed source coding

Convolutional codes

The idea behind a convolutional code is to make every codeword symbol be the weighted sum of the various input message symbols. This is like convolution used in LTI systems to find the output of a system, when you know the input and impulse response.

So we generally find the output of the system convolutional encoder, which is the convolution of the input bit, against the states of the convolution encoder, registers.

Fundamentally, convolutional codes do not offer more protection against noise than an equivalent block code. In many cases, they generally offer greater simplicity of implementation over a block code of equal power. The encoder is usually a simple circuit which has state memory and some feedback logic, normally XOR gates. The decoder can be implemented in software or firmware.

The Viterbi algorithm is the optimum algorithm used to decode convolutional codes. There are simplifications to reduce the computational load. They rely on searching only

the most likely paths. Although not optimum, they have generally found to give good results in the lower noise environments.

Convolutional codes are used in voiceband modems (V.32, V.17, V.34) and in GSM mobile phones, as well as satellite and military communication devices.

Other applications of coding theory

Another concern of coding theory is designing codes that help synchronization. A code may be designed so that a phase shift can be easily detected and corrected and that multiple signals can be sent on the same channel.

Another application of codes, used in some mobile phone systems, is code-division multiple access (CDMA). Each phone is assigned a code sequence that is approximately uncorrelated with the codes of other phones. When transmitting, the code word is used to modulate the data bits representing the voice message. At the receiver, a demodulation process is performed to recover the data. The properties of this class of codes allow many users (with different codes) to use the same radio channel at the same time. To the receiver, the signals of other users will appear to the demodulator only as a low-level noise.

Another general class of codes are the automatic repeat-request (ARQ) codes. In these codes the sender adds redundancy to each message for error checking, usually by adding check bits. If the check bits are not consistent with the rest of the message when it arrives, the receiver will ask the sender to retransmit the message. All but the simplest wide area network protocols use ARQ. Common protocols include SDLC (IBM), TCP (Internet), X.25 (International) and many others. There is an extensive field of research on this topic because of the problem of matching a rejected packet against a new packet. Is it a new one or is it a retransmission? Typically numbering schemes are used, as in TCP."RFC793". *RFCs*. Internet Engineering Task Force (IETF).

Group Testing

Group testing uses codes in a different way. Consider a large group of items in which a very few are different in a particular way (for eg. Defective products or infected test subjects). The idea of group testing is to determine which items are "different" by using as few tests as possible. The origin of the problem has its roots in the Second World War when the United States Army Air Forces needed to test its soldiers for Syphilis. It originated from a ground-breaking paper by Robert Dorfman.

Analog coding

Information is encoded analogously in the neural networks of brains, in analog signal processing, and analog electronics. Aspects of analog coding include analog error correction, analog data compression. analog encryption

Neural coding

Neural coding is a neuroscience-related field concerned with how sensory and other information is represented in the brain by networks of neurons. The main goal of studying neural coding is to characterize the relationship between the stimulus and the individual or ensemble neuronal responses and the relationship among electrical activity of the neurons in the ensemble. It is thought that neurons can encode both digital and analog information, and that neurons follow the principles of information theory and compress information, and detect and correct errors in the signals that are sent throughout the brain and wider nervous system.

Chapter-11

Convolutional Code, Concatenated Error Correction Code and Hadamard Code

Convolutional code

In telecommunication, a **convolutional code** is a type of error-correcting code in which

- each m-bit information symbol (each m-bit string) to be encoded is transformed into an n-bit symbol, where m/n is the code *rate* ($n \geq m$) and
- the transformation is a function of the last k information symbols, where k is the constraint length of the code.

Where convolutional codes are used

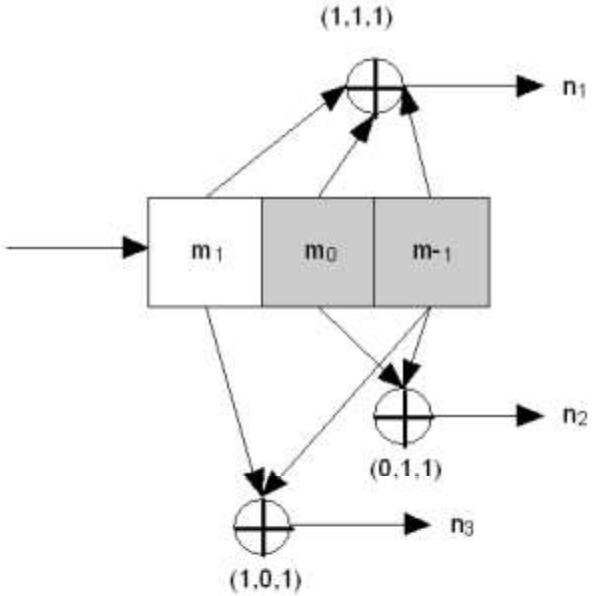
Convolutional codes are used extensively in numerous applications in order to achieve reliable data transfer, including digital video, radio, mobile communication, and satellite communication. These codes are often implemented in concatenation with a hard-decision code, particularly Reed Solomon. Prior to turbo codes, such constructions were the most efficient, coming closest to the Shannon limit.

Convolutional encoding

To convolutionally encode data, start with k memory registers, each holding 1 input bit. Unless otherwise specified, all memory registers start with a value of 0. The encoder has n modulo-2 adders (a modulo 2 adder can be implemented with a single Boolean XOR gate, where the logic is: 0+0 = 0, 0+1 = 1, 1+0 = 1, 1+1 = 0), and n generator polynomials — one for each adder (see figure below). An input bit m_1 is fed into the leftmost register. Using the generator polynomials and the existing values in the remaining registers, the encoder outputs n bits. Now bit shift all register values to the right (m_1 moves to m_0, m_0 moves to m_{-1}) and wait for the next input bit. If there are no remaining input bits, the encoder continues output until all registers have returned to the zero state.

The figure below is a rate 1/3 (m/n) encoder with constraint length (k) of 3. Generator polynomials are $G_1 = (1,1,1)$, $G_2 = (0,1,1)$, and $G_3 = (1,0,1)$. Therefore, output bits are calculated (modulo 2) as follows:

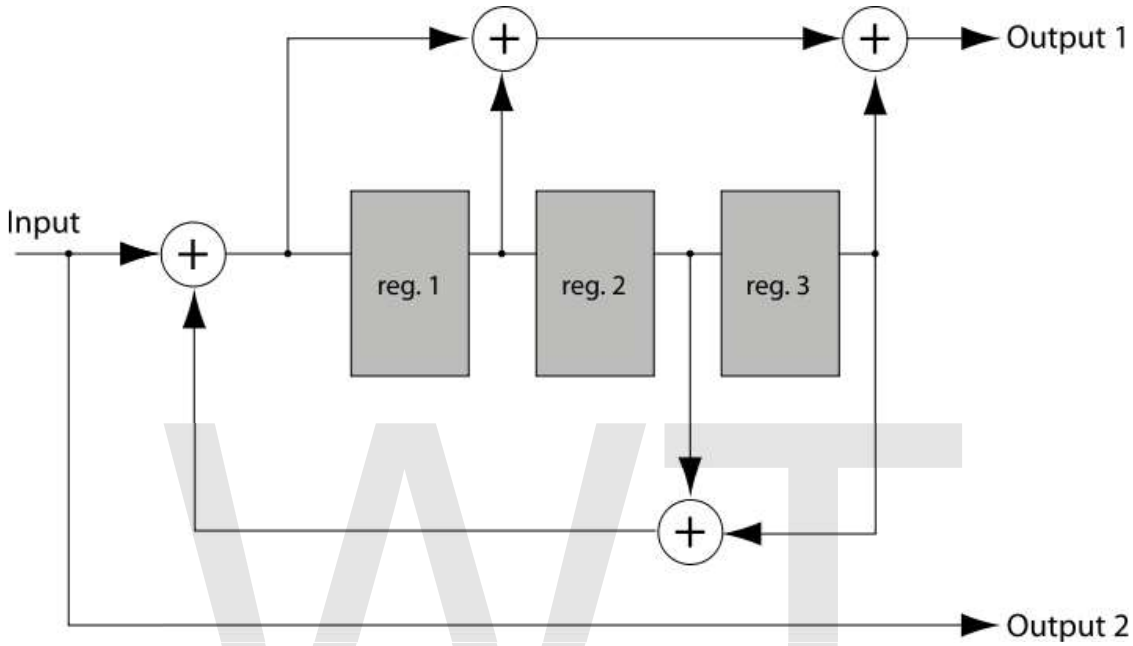
$n_1 = m_1 + m_0 + m_{-1}$
$n_2 = m_0 + m_{-1}$
$n_3 = m_1 + m_{-1}$.



Img.1. Rate 1/3 non-recursive, non-systematic convolutional encoder with constraint length 3

Recursive and non-recursive codes

The encoder on the picture above is a *non-recursive* encoder. Here's an example of a recursive one:



Img.2. Rate 1/2 recursive, systematic convolutional encoder with constraint length 4

One can see that the input being encoded is included in the output sequence too (look at the output 2). Such codes are referred to as *systematic*; otherwise the code is called *non-systematic*.

Recursive codes are almost always systematic and, conversely, non-recursive codes are non-systematic. It isn't a strict requirement, but a common practice.

Impulse response, transfer function, and constraint length

A convolutional encoder is called so because it performs a *convolution* of the input stream with the encoder's *impulse responses*:

$$y_i^j = \sum_{k=0}^{\infty} h_k^j x_{i-k},$$

where x is an input sequence, y^j is a sequence from output j and h^j is an impulse response for output j.

A convolutional encoder is a discrete linear time-invariant system. Every output of an encoder can be described by its own transfer function, which is closely related to a generator polynomial. An impulse response is connected with a transfer function through Z-transform.

Transfer functions for the first (non-recursive) encoder are:

- $H_1(z) = 1 + z^{-1} + z^{-2}$,
- $H_2(z) = z^{-1} + z^{-2}$,
- $H_3(z) = 1 + z^{-2}$.

Transfer functions for the second (recursive) encoder are:

- $H_1(z) = \frac{1 + z^{-1} + z^{-3}}{1 - z^{-2} - z^{-3}}$,
- $H_2(z) = 1$.

Define $m$by

$$m = \max_i polydeg(H_i(1/z))$$

where, for any rational function $f(z) = P(z)/Q(z)$,

$$polydeg(f) = \max(deg(P), deg(Q))$$.

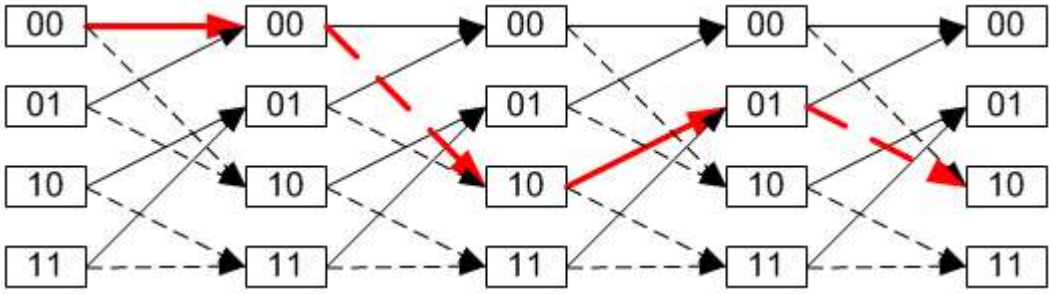
Then $m$is the maximum of the polynomial degrees of the $H_i(1/z)$, and the *constraint length* is defined as $K = m + 1$. For instance, in the first example the constraint length is 3, and in the second the constraint length is 4.

Trellis diagram

A convolutional encoder is a finite state machine. An encoder with *n* binary cells will have 2^n states.

Imagine that the encoder (shown on Img.1, above) has '1' in the left memory cell (m_0), and '0' in the right one (m_{-1}). (m_1 is not really a memory cell because it represents a current value). We will designate such a state as "10". According to an input bit the encoder at the next turn can convert either to the "01" state or the "11" state. One can see that not all transitions are possible (e.g., a decoder can't convert from "10" state to "00" or even stay in "10" state).

All possible transitions can be shown as below:



Img.3. A trellis diagram for the encoder on Img.1. A path through the trellis is shown as a red line. The solid lines indicate transitions where a "0" is input and the dashed lines where a "1" is input.

An actual encoded sequence can be represented as a path on this graph. One valid path is shown in red as an example.

This diagram gives us an idea about *decoding*: if a received sequence doesn't fit this graph, then it was received with errors, and we must choose the nearest *correct* (fitting the graph) sequence. The real decoding algorithms exploit this idea.

Free distance and error distribution

The **free distance** (*d*) is the minimal Hamming distance between different encoded sequences. The *correcting capability* (*t*) of a convolutional code is the number of errors that can be corrected by the code. It can be calculated as

$$t = \left\lfloor \frac{d-1}{2} \right\rfloor .$$

Since a convolutional code doesn't use blocks, processing instead a continuous bitstream, the value of *t* applies to a quantity of errors located relatively near to each other. That is, multiple groups of *t* errors can usually be fixed when they are relatively far apart.

Free distance can be interpreted as the minimal length of an erroneous "burst" at the output of a convolutional decoder. The fact that errors appear as "bursts" should be accounted for when designing a concatenated code with an inner convolutional code. The popular solution for this problem is to interleave data before convolutional encoding, so that the outer block (usually Reed-Solomon) code can correct most of the errors.

Decoding convolutional codes

Several algorithms exist for decoding convolutional codes. For relatively small values of *k*, the Viterbi algorithm is universally used as it provides maximum likelihood

performance and is highly parallelizable. Viterbi decoders are thus easy to implement in VLSI hardware and in software on CPUs with SIMD instruction sets.

Longer constraint length codes are more practically decoded with any of several sequential decoding algorithms, of which the Fano algorithm is the best known. Unlike Viterbi decoding, sequential decoding is not maximum likelihood but its complexity increases only slightly with constraint length, allowing the use of strong, long-constraint-length codes. Such codes were used in the Pioneer program of the early 1970s to Jupiter and Saturn, but gave way to shorter, Viterbi-decoded codes, usually concatenated with large Reed-Solomon error correction codes that steepen the overall bit-error-rate curve and produce extremely low residual undetected error rates.

Both Viterbi and sequential decoding algorithms return hard-decisions: the bits that form the most likely codeword. An approximate confidence measure can be added to each bit by use of the Soft output Viterbi algorithm. Maximum a posteriori (MAP) soft-decisions for each bit can be obtained by use of the BCJR algorithm.

Popular convolutional codes

An especially popular Viterbi-decoded convolutional code, used at least since the Voyager program has a constraint length k of 7 and a rate r of 1/2.

- Longer constraint lengths produce more powerful codes, but the complexity of the Viterbi algorithm increases exponentially with constraint lengths, limiting these more powerful codes to deep space missions where the extra performance is easily worth the increased decoder complexity.

- Mars Pathfinder, Mars Exploration Rover and the Cassini probe to Saturn use a k of 15 and a rate of 1/6; this code performs about 2 dB better than the simpler k=7 code at a cost of 256× in decoding complexity (compared to Voyager mission codes).

Punctured convolutional codes

Puncturing is a technique used to make a m/n rate code from a "basic" rate 1/2 code. It is reached by deletion of some bits in the encoder output. Bits are deleted according to *puncturing matrix*. The following puncturing matrices are the most frequently used:

Code rate	Puncturing matrix	Free distance (for NASA standard K=7 convolutional code)
1/2 (No perf.)	1 1	10
2/3	1 0 1 1	6

3/4	1 0 1 1 1 0	5
5/6	1 0 1 0 1 1 1 0 1 0	4
7/8	1 0 0 0 1 0 1 1 1 1 1 0 1 0	3

For example, if we want to make a code with rate 2/3 using the appropriate matrix from the above table, we should take a basic encoder output and transmit every second bit from the first branch and every bit from the second one. The specific order of transmission is defined by the respective communication standard.

Punctured convolutional codes are widely used in the satellite communications, for example, in INTELSAT systems and Digital Video Broadcasting.

Punctured convolutional codes are also called "perforated".

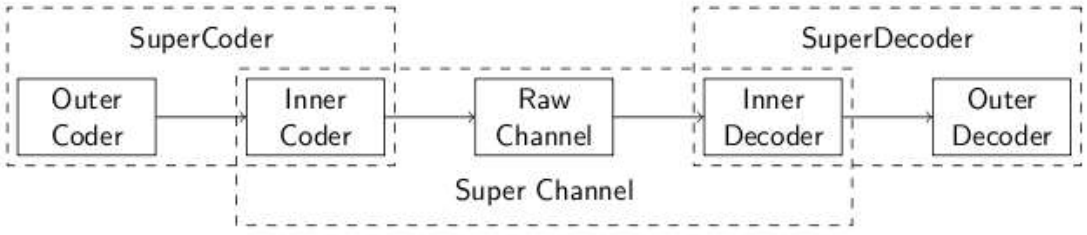
Turbo codes: replacing convolutional codes

Simple Viterbi-decoded convolutional codes are now giving way to turbo codes, a new class of iterated short convolutional codes that closely approach the theoretical limits imposed by Shannon's theorem with much less decoding complexity than the Viterbi algorithm on the long convolutional codes that would be required for the same performance. Concatenation with an outer algebraic code (e.g., Reed-Solomon) addresses the issue of error floors inherent to turbo code designs.

Concatenated error correction code

In coding theory, **concatenated codes** form a class of error-correcting codes that are derived by combining an **inner code** and an **outer code**. They were conceived in 1966 by Dave Forney as a solution for the problem of finding a code that has both exponentially decreasing error probability with increasing block length and polynomial-time decoding complexity.

Description



Let C be a code with length N and rate R over an alphabet A with $K=N*R$ symbols. Let I be another code with length n and rate r over an alphabet B with $k=n*r$ symbols.

The inner code I takes one of k possible inputs, encodes onto an n-tuple from B, transmits, and decodes into one of k possible outputs. We regard this as a (super) channel which can transmit one symbol from the alphabet A, also of size k. We use this channel N times to transmit each of the N symbols in a codeword of C. The *concatenation* of C (as outer code) with I (as inner code) is thus a code of length Nn over the alphabet B.

The *key insight* in this approach is that if I is decoded using a maximum-likelihood approach (thus showing an exponentially decreasing error probability with increasing length), and C is a code with length $N=2^{nr}$ that can be decoded in polynomial time of N, then the concatenated code can be decoded in polynomial time of its combined length $n2^{nr}$ and shows an exponentially decreasing error probability, even if I has exponential decoding complexity.

In a generalization of above concatenation, there are N possible inner codes I_i and the i-th symbol in a codeword of C is transmitted across the inner channel using the i-th inner code. The Justesen codes are examples of generalized concatenated codes, where the outer code is a Reed-Solomon code.

Applications

Concatenated codes were starting to be regularly used for deep space communication with the Voyager program, which launched their first probe in 1977. (A simple concatenation scheme however was already implemented for the 1971 Mariner Mars orbiter mission.) Since then, concatenated codes became the workhorse for efficient error correction coding, and stayed so at least until the invention of turbo codes and LDPC codes.

Typically, the inner code is not a block code but a soft-decision convolutional Viterbi-decoded code with a short constraint length. For the outer code, a longer hard-decision block code, frequently Reed Solomon with 8-bit symbols, is selected. The larger symbol size makes the outer code more robust to burst errors that may occur due to channel

impairments, and because erroneous output of the convolutional code itself is bursty. An interleaving layer is usually added between the two codes to spread burst errors across a wider range.

The combination of an inner Viterbi convolutional code with an outer Reed-Solomon code (known as an RSV code) was first used on Voyager 2, and became a popular construction both within and outside of the space sector. It is still notably used today for satellite communication, such as the DVB-S digital television broadcast standard.

In a more loose sense, any (serial) combination of two or more codes may be referred to as a concatenated code. For example, within the DVB-S2 standard, a highly efficient LDPC code is combined with an algebraic outer code in order to remove any resilient errors left over from the inner LDPC code due to its inherent error floor.

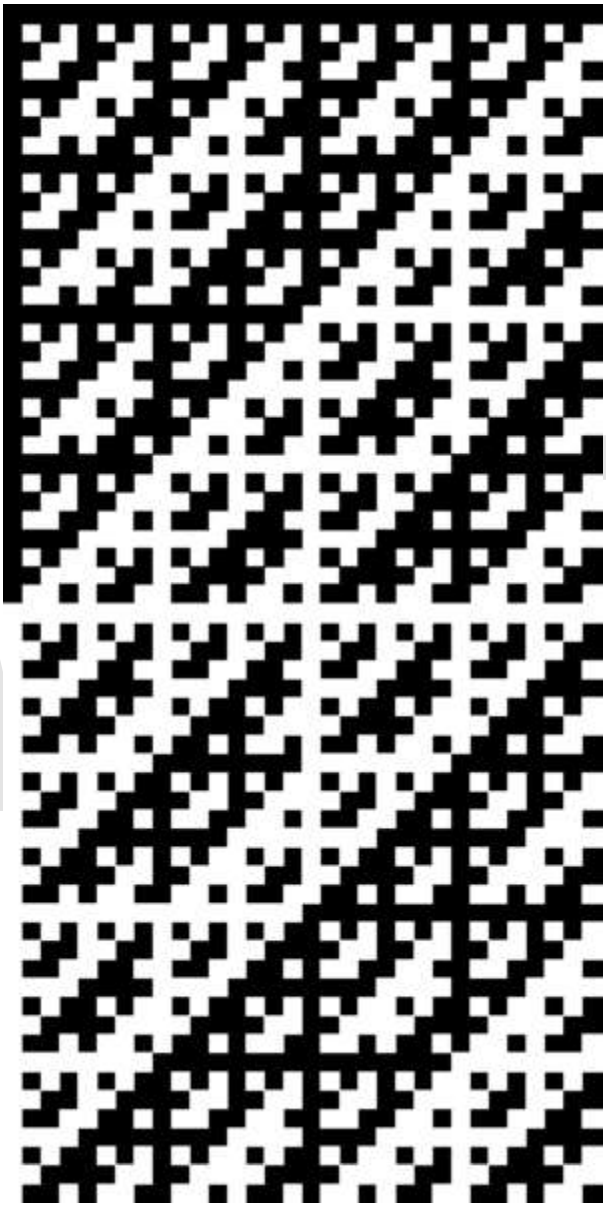
A simple concatenation scheme is also used on the Compact Disc, where an interleaving layer between two Reed-Solomon codes of different sizes effectively spreads errors across different blocks.

Turbo codes: A parallel concatenation approach

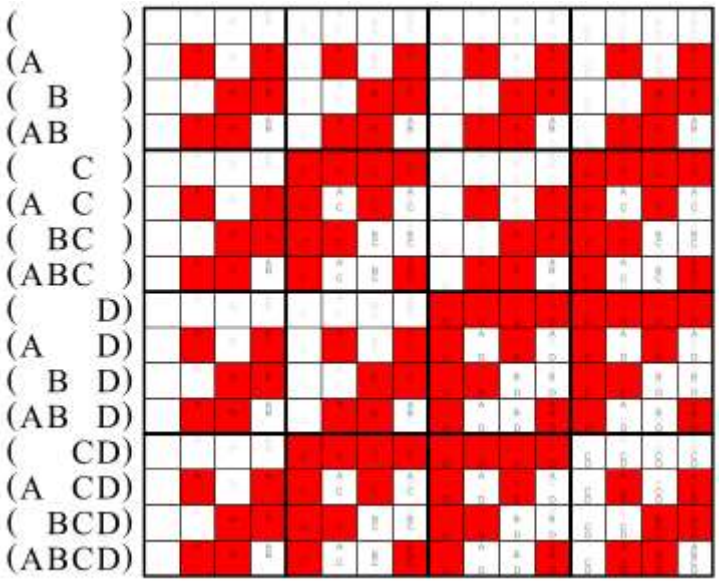
The description above is given for what is now called a serially concatenated code. Turbo codes, as described first in 1993, implemented a parallel concatenation of two convolutional codes, with an interleaver between the two codes and an iterative decoder that would pass information forth and back between the codes. This construction had much higher performance than all previously conceived concatenated codes.

However, a key aspect of turbo codes is their iterated decoding approach. Iterated decoding is now also applied to serial concatenations in order to achieve higher coding gains, such as within serially concatenated convolutional codes (SCCCs). An early form of iterated decoding was notably implemented with 2 to 5 iterations in the "Galileo code" of the Galileo spacecraft.

Hadamard code



Matrix of the Hadamard code (32, 6, 16) for the Reed-Muller Code (1, 5) of the NASA space probe Mariner 9



XOR operations Here the white fields stand for 0 and the red fields for 1

The **Hadamard code**, named after Jacques Hadamard, is a system used for signal error detection and correction. It is one of the family of $[2^n, n + 1, 2^{n-1}]$ codes. Especially for large n it has a poor rate but it is capable of correcting many errors. Hadamard codes can be considered as a special case of Reed–Muller codes. In particular, first order Reed–Muller code are equivalent to Hadamard codes.

Construction

The code is based on Hadamard matrices. If H is a Hadamard matrix of order 2^n the codewords are constructed by taking the rows of H and $-H$ as codewords, where each -1 is replaced by 0. In this way 2^{n+1} code words are constructed that each have a length of 2^n. Since the rows of a Hadamard matrix are orthogonal the minimum distance is 2^{n-1}. In this way a $[2^n, n + 1, 2^{n-1}]$ code is constructed.

The code can also be constructed by creating the parity-check matrix which consists of all 2^{n-1} vectors containing an odd number of 1s, or by using a recursive encoding process.

Decoding

The code has minimum distance 2^{n-1} and hence can correct at most $t = 2^{n-2} - 1$ errors. The algorithm below achieves this.

When a code word is received it is first transformed to a $1/-1$ vector v by changing all 0s to -1. Compute $(v H^{\mathrm{T}})$. The entry with the maximum absolute value corresponds to the row taken as a code word. If it is positive the code word came from H, if it is negative the code word came from $-H$.

Proof: If there were no errors the product ($v\,H^{\mathrm{T}}$) would consist of only zero's and one entry of +/−2^n. If there are errors in v then, in absolute value, some of the zeros become larger and the maximum becomes smaller. Each error that occurs can change a zero with 2. So the zeros can become at most $0 + 2t = 2^{n-1} - 2$. The maximum can decrease at most to $2^n - 2t = 2^n - 2^{n-1} + 2 = 2^{n-1} + 2$. So the maximum that points to the correct row will always be larger in absolute value than the other values in the row.

History

A Hadamard code was used during the 1971 Mariner 9 mission to correct for picture transmission errors. The data words used during this mission were 6 bits long, which represented 64 grayscale values. Because of limitations of the quality of the alignment of the transmitter the maximum useful data length was about 30 bits. Instead of using a repetition code, a [32, 6, 16] Hadamard code was used. Errors of up to 7 bits per word could be corrected using this scheme. Compared to a 5-repetition code, the error correcting properties of this Hadamard code are much better, yet its rate is comparable. The efficient decoding algorithm was an important factor in the decision to use this code. The circuitry used was called the "Green Machine". It employed the fast Fourier transform which can increase the decoding speed by a factor of 3.

Optimality

For value of $n \leq 6$ the Hadamard codes have been proven optimal in the sense of minimum distance.

Chapter-12

Check Digit and Coding Gain

Check digit

A **check digit** is a form of redundancy check used for error detection, the decimal equivalent of a binary checksum. It consists of a single digit computed from the other digits in the message.

With a check digit, one can detect simple errors in the input of a series of digits, such as a single mistyped digit or some permutations of two successive digits.

Design

Check digit algorithms are generally designed to capture *human* transcription errors. In order of complexity, these include the following:

- single digit errors, such as 1 → 2
- transposition errors, such as 12 → 21
- twin errors, such as 11 → 22
- jump transpositions errors, such as 132 → 231
- jump twin errors, such as 131 → 232
- phonetic errors, such as 60 → 16 ("sixty" to "sixteen")

In choosing a system, a high probability of catching errors is traded off against implementation difficulty; simple check digit systems are easily understood and implemented by humans but do not catch as many errors as complex ones, which require sophisticated programs to implement.

A desirable feature is that left-padding with zeros should not change the check digit. This allows variable length digits to be used and the length to be changed.

If there is a single check digit added to the original number, the system will not always capture *multiple* errors, such as two replacement errors (12 → 34) though, typically,

double errors will be caught 90% of the time (both changes would need to change the output by offsetting amounts).

A very simple check digit method would be to take the sum of all digits (digital sum) modulo 10. This would catch any single-digit error, as such an error would always change the sum, but does not catch any transposition errors (switching two digits) as re-ordering does not change the sum.

A slightly more complex method is to take the weighted sum of the digits, modulo 10, with different weights for each number position.

To illustrate this, for example if the weights for a four digit number were 5, 3, 2, 7 and the number to be coded was 4871, then one would take 5×4 + 3×8 + 2×7 + 7×1 = 65, ie 5 modulo 10, and the check digit would be 5, giving 48715.

Systems with weights of 1, 3, 7, or 9, with the weights on neighboring numbers being different, are widely used: for example, 31 31 weights in UPC codes, 13 13 weights in EAN numbers (GS1 algorithm), and the 371 371 371 weights used in United States bank routing transit numbers. This system detects all single-digit errors and around 90% of transposition errors. 1, 3, 7, and 9 are used because they are coprime to 10, so changing any digit changes the check digit; using a coefficient that is divisible by 2 or 5 would lose information (because $5 \cdot 0 = 5 \cdot 2 = 5 \cdot 4 = 5 \cdot 6 = 5 \cdot 8 = 0 \mod 10$) and thus not catch some single-digit errors. Using different weights on neighboring numbers means that most transpositions change the check digit; however, because all weights differ by an even number, this does not catch transpositions of two digits that differ by 5, (0 and 5, 1 and 6, 2 and 7, 3 and 8, 4 and 9), since the 2 and 5 multiply to yield 10.

The ISBN-10 code instead uses modulo 11, which is prime, and all the number positions have different weights $1, 2, \ldots, 10$. This system thus detects all single digit substitution and transposition errors (including jump transpositions), but at the cost of the check digit possibly being 10, represented by "X". (An alternative is simply to avoid using the serial numbers which result in an "X" check digit.) ISBN-13 instead uses the GS1 algorithm used in EAN numbers.

More complicated algorithms include the Luhn algorithm (1954), which captures 98% of single digit transposition errors (it does not detect 90 ↔ 09), while more sophisticated is the Verhoeff algorithm (1969), which catches all single digit substitution and transposition errors, and many (but not all) more complex errors. Both these methods use a single check digit and will therefore fail to capture around 10% of more complex errors. To reduce this failure rate, it is necessary to use more than one check digit (for example, the modulo 97 check referred to below, which uses two check digits) and/or to use a wider range of characters in the check digit, for example letters plus numbers).

Examples

UPC

The final digit of a Universal Product Code is a check digit computed as follows:

1. Add the digits (up to but not including the check digit) in the odd-numbered positions (first, third, fifth, etc.) together and multiply by three.
2. Add the digits (up to but not including the check digit) in the even-numbered positions (second, fourth, sixth, etc.) to the result.
3. Take the remainder of the result divided by 10 (modulo operation) and subtract this from 10 to derive the check digit.

For instance, the UPC-A barcode for a box of tissues is "036000241457". The last digit is the check digit "7", and if the other numbers are correct then the check digit calculation must produce 7.

1. Add the odd number digits: 0+6+0+2+1+5 = 14
2. Multiply the result by 3: 14 × 3 = 42
3. Add the even number digits: 3+0+0+4+4 = 11
4. Add the two results together: 42 + 11 = 53
5. To calculate the check digit, take the remainder of (53 / 10), which is also known as (53 modulo 10), and subtract from 10. Therefore, the check digit value is 7.

Another example: to calculate the check digit for the following food item "01010101010".

1. Add the odd number digits: 0+0+0+0+0+0 = 0
2. Multiply the result by 3: 0 x 3 = 0
3. Add the even number digits: 1+1+1+1+1 = 5
4. Add the two results together: 0 + 5 = 5
5. To calculate the check digit, take the remainder of (5 / 10), which is also known as (5 modulo 10), and subtract from 10 (i.e. (10 - 5 modulo 10) = 5. Therefore, the check digit value is 5.
6. If the remainder is 0, subtracting from 10 would give 10. In that case, use 0 as the check digit.

ISBN 10

The final character of a ten digit International Standard Book Number is a check digit computed so that multiplying each digit by its position in the number (counting from the right) and taking the sum of these products modulo 11 is 0. The digit the farthest to the right (which is multiplied by 1) is the check digit, chosen to make the sum correct. It may need to have the value 10, which is represented as the letter X. For example, take the ISBN 0-201-53082-1. The sum of products is 0×10 + 2×9 + 0×8 + 1×7 + 5×6 + 3×5 + 0×4 + 8×3 + 2×2 + 1×1 = 99 ≡ 0 modulo 11. So the ISBN is valid.

While this may seem more complicated than the first scheme, it can be validated simply by adding all the products together then dividing by 11. The sum can be computed without any multiplications by initializing two variables, `t` and `sum`, to 0 and repeatedly performing `t = t + digit; sum = sum + t;` (which can be expressed in C as `sum += t += digit;`). If the final `sum` is a multiple of 11, the ISBN is valid.

ISBN 13

ISBN 13 (in use January 2007) is equal to the EAN-13 code found underneath a book's barcode. Its check digit is generated the same way as the UPC except that the even digits are multiplied by 3 instead of the odd digits.

EAN (GLN,GTIN, EAN numbers administered by GS1)

EAN (European Article Number) check digits (administered by GS1) are calculated by summing the even position numbers and multiplying by 3 and then by adding the sum of the odd position numbers. The final digit of the result is subtracted from 10 to calculate the check digit (or left as is if already zero). A GS1 check digit calculator and detailed documentation is online at GS1's website.

Other examples of check digits

- The tenth digit of the National Provider Identifier for the US healthcare industry
- The Australian Tax File Number (based on modulo 11)
- The Guatemalan Tax Number (NIT - Número de Identificación Tributaria) based on modulo 11
- The North American CUSIP number
- The final (ninth) digit of the routing transit number, a bank code used in the United States
- The International SEDOL number
- The International Securities Identifying Number (ISIN)
- The International CAS registry number's final digit.
- Modulo 10 check digits in credit card account numbers, calculated with the Luhn algorithm.
    - Also used in the Norwegian KID (customer identification number) numbers used in bank giros (credit transfer).
- The final character encoded in a magnetic stripe card is a computed Longitudinal redundancy check
- final digit of a POSTNET code

Coding gain

In coding theory and related engineering problems, **coding gain** is the measure in the difference between the signal to noise ratio (SNR) levels between the uncoded system and coded system required to reach the same bit error rate (BER) levels when used with the error correcting code (ECC).

Example

If the uncoded BPSK system in AWGN environment has a Bit error rate (BER) of 10^{-2} at the SNR level 4dB, and the corresponding coded (*e.g.*, BCH) system has the same BER at an SNR level of 2.5dB, then we say the *coding gain* = 4dB-2.5dB = 1.5dB, due to the code used (in this case BCH).

Power-limited regime

In the *power-limited regime* (where the nominal spectral efficiency $\rho \leq 2$[b/2D or b/s/Hz], *i.e.* the domain of binary signaling), the effective coding gain $\gamma_{eff}(A)$ of a signal set A at a given target error probability per bit $P_b(E)$ is defined as the difference in dB between the E_b / N_0 required to achieve the target $P_b(E)$ with A and the E_b / N_0 required to achieve the target $P_b(E)$ with 2-PAM or (2×2)-QAM (*i.e.* no coding). The nominal coding gain $\gamma_c(A)$ is defined as

$$\gamma_c(A) = \frac{d^2_{\min}(A)}{4E_b}.$$

This definition is normalized so that $\gamma_c(A) = 1$ for 2-PAM or (2×2)-QAM. If the average number of nearest neighbors per transmitted bit $K_b(A)$ is equal to one, the effective coding gain $\gamma_{eff}(A)$ is approximately equal to the nominal coding gain $\gamma_c(A)$. However, if $K_b(A) > 1$, the effective coding gain $\gamma_{eff}(A)$ is less than the nominal coding gain $\gamma_c(A)$ by an amount which depends on the steepness of the $P_b(E)$ *vs.* E_b / N_0 curve at the target $P_b(E)$. This curve can be plotted using the union bound estimate (UBE)

$$P_b(E) \approx K_b(A) Q\sqrt{(2\gamma_c(A)E_b/N_0)},$$

where $Q(\cdot)$ denotes the Gaussian probability of error function.

For the special case of a binary linear block code C with parameters (n,k,d), the nominal spectral efficiency is $\rho = 2k / n$ and the nominal coding gain is kd/n.

Example

The table below lists the nominal spectral efficiency, nominal coding gain and effective coding gain at $P_b(E) \approx 10^{-5}$ for Reed-Muller codes of length $n \leq 64$:

Code	ρ	γ_c	γ_c (dB)	K_b	γ_{eff} (dB)
[8,7,2]	1.75	7/4	2.43	4	2.0
[8,4,4]	1.0	2	3.01	4	2.6
[16,15,2]	1.88	15/8	2.73	8	2.1
[16,11,4]	1.38	11/4	4.39	13	3.7
[16,5,8]	0.63	5/2	3.98	6	3.5
[32,31,2]	1.94	31/16	2.87	16	2.1
[32,26,4]	1.63	13/4	5.12	48	4.0
[32,16,8]	1.00	4	6.02	39	4.9
[32,6,16]	0.37	3	4.77	10	4.2
[64,63,2]	1.97	63/32	2.94	32	1.9
[64,57,4]	1.78	57/16	5.52	183	4.0
[64,42,8]	1.31	21/4	7.20	266	5.6
[64,22,16]	0.69	11/2	7.40	118	6.0
[64,7,32]	0.22	7/2	5.44	18	4.6

Bandwidth-limited regime

In the *bandwidth-limited regime* ($\rho > 2b / 2D$, *i.e.* the domain of non-binary signaling), the effective coding gain $\gamma_{eff}(A)$ of a signal set A at a given target error rate $P_s(E)$ is defined as the difference in dB between the SNR_{norm} required to achieve the target $P_s(E)$ with A and the SNR_{norm} required to achieve the target $P_s(E)$ with M-PAM or (M×M)-QAM (*i.e.* no coding). The nominal coding gain $\gamma_c(A)$ is defined as

$$\gamma_c(A) = \frac{(2^\rho - 1) d^2_{\min}(A)}{6E_s}.$$

This definition is normalized so that $\gamma_c(A) = 1$ for M-PAM or (M×M)-QAM. The UBE becomes

$$P_s(E) \approx K_s(A) Q\sqrt{(3\gamma_c(A) SNR_{norm})},$$

where $K_s(A)$ is the average number of nearest neighbors per two dimensions.

Chapter-13

Hamming Code

In telecommunication, a **Hamming code** is a linear error-correcting code named after its inventor, Richard Hamming. Hamming codes can detect up to two simultaneous bit errors, and correct single-bit errors; thus, reliable communication is possible when the Hamming distance between the transmitted and received bit patterns is less than or equal to one. By contrast, the simple parity code cannot correct errors, and can only detect an odd number of errors.

In mathematical terms, Hamming codes are a class of binary linear codes. For each integer $m \geq 2$ there is a code with *m* parity bits and $2^m - m - 1$ data bits. The parity-check matrix of a Hamming code is constructed by listing all columns of length *m* that are pairwise independent. Hamming codes are an example of perfect codes, codes that exactly match the theoretical upper bound on the number of distinct code words for a given number of bits and ability to correct errors.

Because of the simplicity of Hamming codes, they are widely used in computer memory (RAM). In particular, a single-error-correcting *and* double-error-detecting variant commonly referred to as **SECDED**.

History

Hamming worked at Bell Labs in the 1940s on the Bell Model V computer, an electromechanical relay-based machine with cycle times in seconds. Input was fed in on punched cards, which would invariably have read errors. During weekdays, special code would find errors and flash lights so the operators could correct the problem. During after-hours periods and on weekends, when there were no operators, the machine simply moved on to the next job.

Hamming worked on weekends, and grew increasingly frustrated with having to restart his programs from scratch due to the unreliability of the card reader. Over the next few years he worked on the problem of error-correction, developing an increasingly powerful array of algorithms. In 1950 he published what is now known as Hamming Code, which remains in use today in applications such as ECC memory.

Codes predating Hamming

A number of simple error-detecting codes were used before Hamming codes, but none were as effective as Hamming codes in the same overhead of space.

Parity

Parity adds a single bit that indicates whether the number of 1 bits in the preceding data was even or odd. If an odd number of bits is changed in transmission, the message will change parity and the error can be detected at this point. (Note that the bit that changed may have been the parity bit itself!) The most common convention is that a parity value of **1** indicates that there is an **odd** number of ones in the data, and a parity value of **0** indicates that there is an **even** number of ones in the data. In other words: the data and the parity bit **together** should contain an even number of 1s.

Parity checking is not very robust, since if the number of bits changed is even, the check bit will be valid and the error will not be detected. Moreover, parity does not indicate which bit contained the error, even when it can detect it. The data must be discarded entirely and re-transmitted from scratch. On a noisy transmission medium, a successful transmission could take a long time or may never occur. However, while the quality of parity checking is poor, since it uses only a single bit, this method results in the least overhead. Furthermore, parity checking does allow for the restoration of an erroneous bit when its position is known.

Two-out-of-five code

A two-out-of-five code is an encoding scheme which uses five digits consisting of exactly three 0s and two 1s. This provides ten possible combinations, enough to represent the digits 0 - 9. This scheme can detect all single bit-errors and all odd numbered bit-errors. However it still cannot correct for these errors.

Repetition

Another code in use at the time repeated every data bit several times in order to ensure that it got through. For instance, if the data bit to be sent was a 1, an n=3 *repetition code* would send "111". If the three bits received were not identical, an error occurred. If the channel is clean enough, most of the time only one bit will change in each triple. Therefore, 001, 010, and 100 each correspond to a 0 bit, while 110, 101, and 011 correspond to a 1 bit, as though the bits counted as "votes" towards what the original bit was. A code with this ability to reconstruct the original message in the presence of errors is known as an *error-correcting* code. This triple repetition code is actually the simplest Hamming code with $m = 2$, since there are 2 parity bits, and $2^2 - 2 - 1 = 1$ data bit.

Such codes cannot correctly repair all errors, however. In our example, if the channel flipped two bits and the receiver got "001", the system would detect the error, but conclude that the original bit was 0, which is incorrect. If we increase the number of

times we duplicate each bit to four, we can detect all two-bit errors but can't correct them (the votes "tie"); at five, we can correct all two-bit errors, but not all three-bit errors.

Moreover, the repetition code is extremely inefficient, reducing throughput by three times in our original case, and the efficiency drops drastically as we increase the number of times each bit is duplicated in order to detect and correct more errors.

Hamming codes

If more error-correcting bits are included with a message, and if those bits can be arranged such that different incorrect bits produce different error results, then bad bits could be identified. In a 7-bit message, there are seven possible single bit errors, so three error control bits could potentially specify not only that an error occurred but also which bit caused the error.

Hamming studied the existing coding schemes, including two-of-five, and generalized their concepts. To start with, he developed a nomenclature to describe the system, including the number of data bits and error-correction bits in a block. For instance, parity includes a single bit for any data word, so assuming ASCII words with 7-bits, Hamming described this as an *(8,7)* code, with eight bits in total, of which 7 are data. The repetition example would be *(3,1)*, following the same logic. The code rate is the second number divided by the first, for our repetition example, 1/3.

Hamming also noticed the problems with flipping two or more bits, and described this as the "distance" (it is now called the *Hamming distance*, after him). Parity has a distance of 2, as any two bit flips will be invisible. The (3,1) repetition has a distance of 3, as three bits need to be flipped in the same triple to obtain another code word with no visible errors. A (4,1) repetition (each bit is repeated four times) has a distance of 4, so flipping two bits can be detected, but not corrected. When three bits flip in the same group there can be situations where the code corrects towards the wrong code word.

Hamming was interested in two problems at once; increasing the distance as much as possible, while at the same time increasing the code rate as much as possible. During the 1940s he developed several encoding schemes that were dramatic improvements on existing codes. The key to all of his systems was to have the parity bits overlap, such that they managed to check each other as well as the data.

General algorithm

The following general algorithm generates a single-error correcting (SEC) code for any number of bits.

1. Number the bits starting from 1: bit 1, 2, 3, 4, 5, etc.
2. Write the bit numbers in binary. 1, 10, 11, 100, 101, etc.
3. All bit positions that are powers of two (have only one 1 bit in the binary form of their position) are parity bits.

4. All other bit positions, with two or more 1 bits in the binary form of their position, are data bits.
5. Each data bit is included in a unique set of 2 or more parity bits, as determined by the binary form of its bit position.
    1. Parity bit 1 covers all bit positions which have the least significant bit set: bit 1 (the parity bit itself), 3, 5, 7, 9, etc.
    2. Parity bit 2 covers all bit positions which have the second least significant bit set: bit 2 (the parity bit itself), 3, 6, 7, 10, 11, etc.
    3. Parity bit 4 covers all bit positions which have the third least significant bit set: bits 4–7, 12–15, 20–23, etc.
    4. Parity bit 8 covers all bit positions which have the fourth least significant bit set: bits 8–15, 24–31, 40–47, etc.
    5. In general each parity bit covers all bits where the binary AND of the parity position and the bit position is non-zero.

The form of the parity is irrelevant. Even parity is simpler from the perspective of theoretical mathematics, but there is no difference in practice.

This general rule can be shown visually:

	Bit position	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	
	Encoded data bits	**p1**	**p2**	**d1**	**p4**	**d2**	**d3**	**d4**	**p8**	**d5**	**d6**	**d7**	**d8**	**d9**	**d10**	**d11**	**p16**	**d12**	**d13**	**d14**	**d15**	
Parity bit coverage	**p1**	X		X		X		X		X		X		X		X		X		X		
	p2		X	X			X	X			X	X			X	X			X	X		...
	p4				X	X	X	X					X	X	X	X					X	
	p8								X	X	X	X	X	X	X	X						
	p16																X	X	X	X	X	

Shown are only 20 encoded bits (5 parity, 15 data) but the pattern continues indefinitely. The key thing about Hamming Codes that can be seen from visual inspection is that any given bit is included in a unique set of parity bits. To check for errors, check all of the parity bits. The pattern of errors, called the error syndrome, identifies the bit in error. If all parity bits are correct, there is no error. Otherwise, the sum of the positions of the erroneous parity bits identifies the erroneous bit. For example, if the parity bits in positions 1, 2 and 8 indicate an error, then bit 1+2+8=11 is in error. If only one parity bit indicates an error, the parity bit itself is in error.

If, in addition, an overall parity bit (bit 0) is included, the code can detect (but not correct) any two-bit error, making a SECDED code. The overall parity indicates whether the total number of errors is even or odd. If the basic Hamming code detects an error, but the overall parity says that there are an even number of errors, an uncorrectable 2-bit error has occurred.

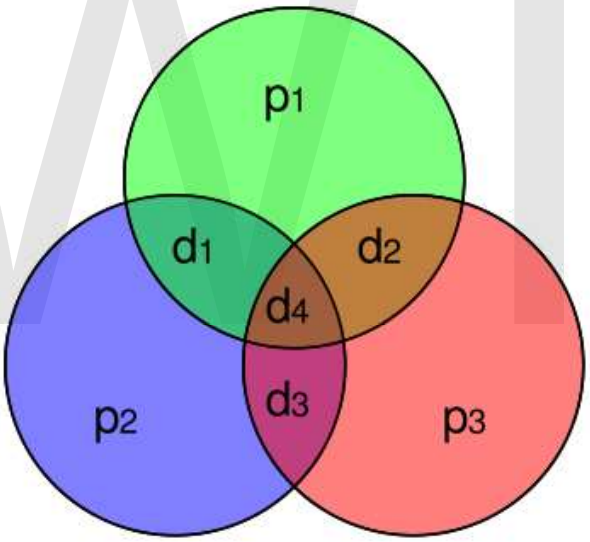
Hamming codes with additional parity (SECDED)

These codes have a minimum distance of 3, which means that the code can detect and correct a single error, but a double bit error is indistinguishable from a different code with a single bit error. Thus, they can detect double-bit errors only if correction is not attempted.

By including an extra parity bit, it is possible to increase the minimum distance of the Hamming code to 4. This gives the code the ability to detect and correct a single error and at the same time detect (but not correct) a double error. (It could also be used to detect up to 3 errors but not correct any.)

This code system is popular in computer memory systems, where it is known as SECDED ("single error correction, double error detection"). Particularly popular is the (72,64) code, a truncated (127,120) Hamming code plus an additional parity bit, which has the same space overhead as a (9,8) parity code.

Hamming(7,4) code



Graphical depiction of the 4 data bits and 3 parity bits and which parity bits apply to which data bits

In 1950, Hamming introduced the (7,4) code. It encodes 4 data bits into 7 bits by adding three parity bits. Hamming(7,4) can detect and correct single-bit errors. With the addition of an overall parity bit, it can also detect (but not correct) double-bit errors.

Construction of G and H

The matrix $\mathbf{G} := \begin{pmatrix} I_k | & -A^T \end{pmatrix}$ is called a (Canonical) generator matrix of a linear (n,k) code,

and $\mathbf{H} := \left(A|I_{n-k}\right)$ is called a parity-check matrix.

This is the construction of **G** and **H** in standard (or systematic) form. Regardless of form, **G** and **H** for linear block codes must satisfy

$\mathbf{H}\,\mathbf{G}^T = \mathbf{0}$, an all-zeros matrix [Moon, p. 89].

Since (7,4,3)=(n,k,d)=[$2^m - 1$, 2^m−1-m, m]. The parity-check matrix **H** of a Hamming code is constructed by listing all columns of length m that are pair-wise independent.

Thus **H** is a matrix whose left side is all of the nonzero n-tuples where order of the n-tuples in the columns of matrix does not matter. The right hand side is just the (n-k)-identity matrix.

So **G** can be obtained from **H** by taking the transpose of the left hand side of **H** with the identity k-identity matrix on the left hand side of **G**.

The code generator matrix $\mathbf{G}$ and the parity-check matrix $\mathbf{H}$ are:

$$\mathbf{H} := \begin{pmatrix} 1 & 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 1 \end{pmatrix}_{3,7}$$

and

$$\mathbf{G} := \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}_{4,7}.$$

Finally, these matrices can be mutated into equivalent non-systematic codes by the following operations [Moon, p. 85]:

- Column permutations (swapping columns)
- Elementary row operations (replacing a row with a linear combination of rows)

Encoding

Example

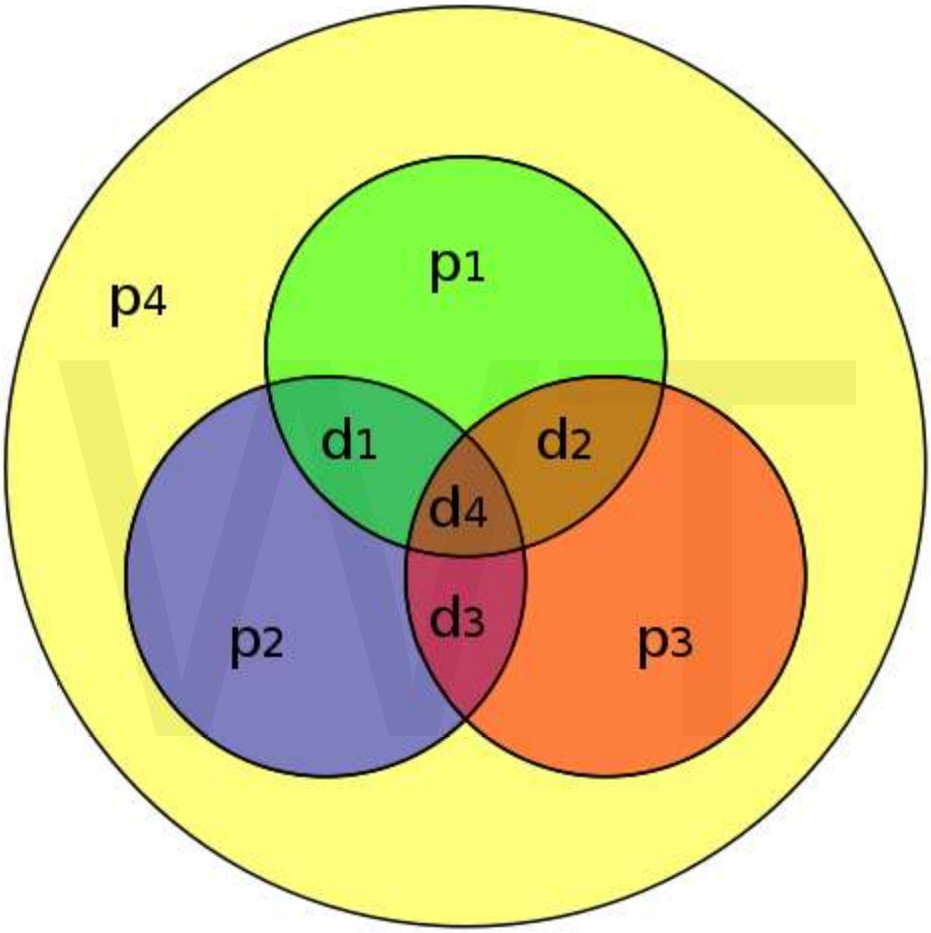
From the above matrix we have 2^k=2^4=16 codewords. The codewords $\overrightarrow{x}$ of this binary code can be obtained from $\overrightarrow{x} = \overrightarrow{a}G$. With $\overrightarrow{a} = a_1 a_2 a_3 a_4$ with a_i exist in F_2 (A field with two elements namely 0 and 1).

Thus the codewords are all the 4-tuples (k-tuples).

Therefore,

(1,0,1,1) gets encoded as (1,0,1,1,0,1,0).

Hamming(7,4) code with an additional parity bit



The same (7,4) example from above with an extra parity bit

The Hamming(7,4) can easily be extended to an (8,4) code by adding an extra parity bit on top of the (7,4) encoded word. This can be summed up with the revised matrices:

$$\mathbf{G} := \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \end{pmatrix}_{4,8}$$

and

$$\mathbf{H} := \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}_{4,8}.$$

Note that H is not in standard form. To obtain G, elementary row operations can be used to obtain an equivalent matrix to H in systematic form:

$$\mathbf{H} = \left(\begin{array}{cccc|cccc} 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \end{array}\right)_{4,8}.$$

For example, the first row in this matrix is the sum of the second and third rows of H in non-systematic form. Using the systematic construction for Hamming codes from above, the matrix A is apparent and the systematic form of G is written as

$$\mathbf{G} = \left(\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \end{array}\right)_{4,8}.$$

The non-systematic form of G can be row reduced (using elementary row operations) to match this matrix.

The addition of the fourth row effectively computes the sum of all the codeword bits (data and parity) as the fourth parity bit.

For example, 1011 is encoded into 01100110 where blue digits are data; red digits are parity from the Hamming(7,4) code; and the green digit is the parity added by Hamming(8,4). The green digit makes the parity of the (7,4) code even.

Finally, it can be shown that the minimum distance has increased from 3, as with the (7,4) code, to 4 with the (8,4) code. Therefore, the code can be defined as Hamming(8,4,4).

Chapter-14

Forward Error Correction and EXIT Chart

Forward error correction

In telecommunication and information theory, **forward error correction (FEC)** (also called **channel coding**) is a system of error control for data transmission, whereby the sender adds systematically generated redundant data to its messages, also known as an **error-correcting code** (ECC). The American mathematician Richard Hamming pioneered this field in the 1940s and invented the first FEC code, the Hamming (7,4) code, in 1950.

The carefully designed redundancy allows the receiver to detect and correct a limited number of errors occurring anywhere in the message without the need to ask the sender for additional data. FEC gives the receiver an ability to correct errors without needing a reverse channel to request retransmission of data, but this advantage is at the cost of a fixed higher forward channel bandwidth. FEC is therefore applied in situations where retransmissions are relatively costly, or impossible such as when broadcasting to multiple receivers. In particular, FEC information is usually added to mass storage devices to enable recovery of corrupted data.

FEC processing in a receiver may be applied to a digital bit stream or in the demodulation of a digitally modulated carrier. For the latter, FEC is an integral part of the initial analog-to-digital conversion in the receiver. The Viterbi decoder implements a soft-decision algorithm to demodulate digital data from an analog signal corrupted by noise. Many FEC coders can also generate a bit-error rate (BER) signal which can be used as feedback to fine-tune the analog receiving electronics.

The maximum fractions of errors or of missing bits that can be corrected is determined by the design of the FEC code, so different forward error correcting codes are suitable for different conditions.

How it works

FEC is accomplished by adding redundancy to the transmitted information using a predetermined algorithm. A redundant bit may be a complex function of many original information bits. The original information may or may not appear literally in the encoded output; codes that include the unmodified input in the output are **systematic**, while those that do not are **non-systematic**.

A simplistic example of FEC is to transmit each data bit 3 times, which is known as a (3,1) repetition code. Through a noisy channel, a receiver might see 8 versions of the output, see table below.

Triplet received	**Interpreted as**
000	0 (error free)
001	0
010	0
100	0
111	1 (error free)
110	1
101	1
011	1

This allows an error in any one of the three samples to be corrected by "majority vote" or "democratic voting". The correcting ability of this FEC is:

- Up to 1 bit of triplet in error, **or**
- up to 2 bits of triplet omitted (cases not shown in table).

Though simple to implement and widely used, this triple modular redundancy is a relatively inefficient FEC. Better FEC codes typically examine the last several dozen, or even the last several hundred, previously received bits to determine how to decode the current small handful of bits (typically in groups of 2 to 8 bits).

Averaging noise to reduce errors

FEC could be said to work by "averaging noise"; since each data bit affects many transmitted symbols, the corruption of some symbols by noise usually allows the original user data to be extracted from the other, uncorrupted received symbols that also depend on the same user data.

- Because of this "risk-pooling" effect, digital communication systems that use FEC tend to work well above a certain minimum signal-to-noise ratio and not at all below it.

- This *all-or-nothing tendency* — the cliff effect — becomes more pronounced as stronger codes are used that more closely approach the theoretical Shannon limit.
- Interleaving FEC coded data can reduce the all or nothing properties of transmitted FEC codes when the channel errors tend to occur in bursts. However, this method has limits; it is best used on narrowband data.

Most telecommunication systems used a fixed channel code designed to tolerate the expected worst-case bit error rate, and then fail to work at all if the bit error rate is ever worse. However, some systems adapt to the given channel error conditions: hybrid automatic repeat-request uses a fixed FEC method as long as the FEC can handle the error rate, then switches to ARQ when the error rate gets too high; adaptive modulation and coding uses a variety of FEC rates, adding more error-correction bits per packet when there are higher error rates in the channel, or taking them out when they are not needed.

Types of FEC

The two main categories of FEC codes are block codes and convolutional codes.

- Block codes work on fixed-size blocks (packets) of bits or symbols of predetermined size. Practical block codes can generally be decoded in polynomial time to their block length.
- Convolutional codes work on bit or symbol streams of arbitrary length. They are most often decoded with the Viterbi algorithm, though other algorithms are sometimes used. Viterbi decoding allows asymptotically optimal decoding efficiency with increasing constraint length of the convolutional code, but at the expense of exponentially increasing complexity. A convolutional code can be turned into a block code, if desired, by "tail-biting".

There are many types of block codes, but among the classical ones the most notable is Reed-Solomon coding because of its widespread use on the Compact disc, the DVD, and in hard disk drives. Golay, BCH, Multidimensional parity, and Hamming codes are other examples of classical block codes.

Hamming ECC is commonly used to correct NAND flash memory errors. This provides single-bit error correction and 2-bit error detection. Hamming codes are only suitable for more reliable single level cell (SLC) NAND. Denser multi level cell (MLC) NAND requires stronger multi-bit correcting ECC such as BCH or Reed-Solomon.

Classical block codes are usually implemented using **hard-decision** algorithms, which means that for every input and output signal a hard decision is made whether it corresponds to a one or a zero bit. In contrast, **soft-decision** algorithms like the Viterbi decoder process (discretized) analog signals, which allows for much higher error-correction performance than hard-decision decoding.

Nearly all classical block codes apply the algebraic properties of finite fields.

Concatenated FEC codes for improved performance

Classical (algebraic) block codes and convolutional codes are frequently combined in **concatenated** coding schemes in which a short constraint-length Viterbi-decoded convolutional code does most of the work and a block code (usually Reed-Solomon) with larger symbol size and block length "mops up" any errors made by the convolutional decoder.

Concatenated codes have been standard practice in satellite and deep space communications since Voyager 2 first used the technique in its 1986 encounter with Uranus.

Low-density parity-check (LDPC)

Low-density parity-check (LDPC) codes are a class of recently re-discovered highly efficient linear block codes. They can provide performance very close to the channel capacity (the theoretical maximum) using an iterated soft-decision decoding approach, at linear time complexity in terms of their block length. Practical implementations can draw heavily from the use of parallelism.

LDPC codes were first introduced by Robert G. Gallager in his PhD thesis in 1960, but due to the computational effort in implementing en- and decoder and the introduction of Reed-Solomon codes, they were mostly ignored until recently.

LDPC codes are now used in many recent high-speed communication standards, such as DVB-S2 (Digital video broadcasting), WiMAX (IEEE 802.16e standard for microwave communications), High-Speed Wireless LAN (IEEE 802.11n), 10GBase-T Ethernet (802.3an) and G.hn/G.9960 (ITU-T Standard for networking over power lines, phone lines and coaxial cable).

Turbo codes

Turbo coding is an iterated soft-decoding scheme that combines two or more relatively simple convolutional codes and an interleaver to produce a block code that can perform to within a fraction of a decibel of the Shannon limit. Predating LDPC codes in terms of practical application, they now provide similar performance.

One of the earliest commercial applications of turbo coding was the CDMA2000 1x (TIA IS-2000) digital cellular technology developed by Qualcomm and sold by Verizon Wireless, Sprint, and other carriers. It is also used for the evolution of CDMA2000 1x specifically for Internet access, 1xEV-DO (TIA IS-856). Like 1x, EV-DO was developed by Qualcomm, and is sold by Verizon Wireless, Sprint, and other carriers (Verizon's marketing name for 1xEV-DO is *Broadband Access*, Sprint's consumer and business marketing names for 1xEV-DO are *Power Vision* and *Mobile Broadband*, respectively.).

Local decoding and testing of codes

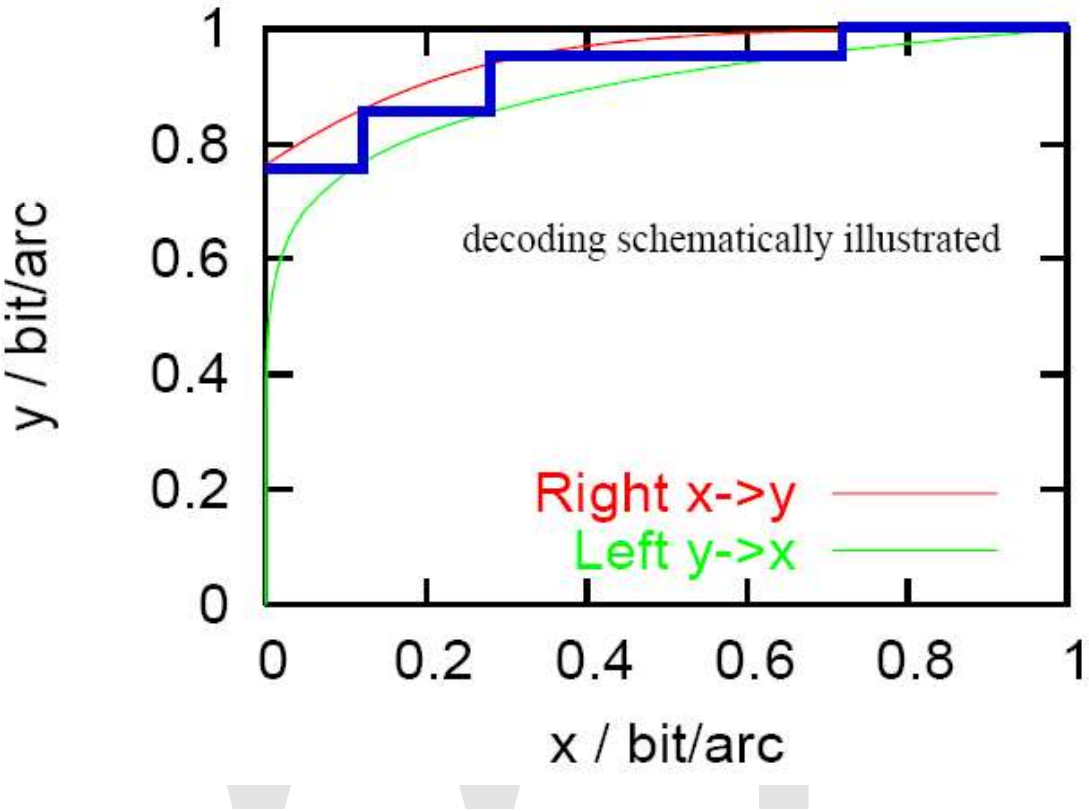
Sometimes it is only necessary to decode single bits of the message, or to check whether a given signal is a codeword, and do so without looking at the entire signal. This can make sense in a streaming setting, where codewords are too large to be classically decoded fast enough and where only a few bits of the message are of interest for now. Also such codes have become an important tool in computational complexity theory, e.g., for the design of probabilistically checkable proofs.

Locally decodable codes are error-correcting codes for which single bits of the message can be probabilistically recovered by only looking at a small (say constant) number of positions of a codeword, even after the codeword has been corrupted at some constant fraction of positions. Locally testable codes are error-correcting codes for which it can be checked probabilistically whether a signal is close to a codeword by only looking at a small number of positions of the signal.

List of error-correcting codes

- BCH code
- Constant-weight code
- Convolutional code
- Group codes
- Golay codes, of which the Binary Golay code is of practical interest
- Goppa code, used in the McEliece cryptosystem
- Hadamard code
- Hagelbarger code
- Hamming code
- Latin square based code for non-white noise (prevalent for example in broadband over powerlines)
- Lexicographic code
- Long code
- Low-density parity-check code, also known as Gallager code, as the archetype for sparse graph codes
- LT code, which is a near-optimal rateless erasure correcting code (Fountain code)
- m of n codes
- Online code, a near-optimal rateless erasure correcting code
- Raptor code, a near-optimal rateless erasure correcting code
- Reed–Solomon error correction
- Reed–Muller code
- Repeat-accumulate code
- Repetition codes, such as Triple modular redundancy
- Tornado code, a near-optimal erasure correcting code, and the precursor to Fountain codes
- Turbo code
- Walsh-Hadamard code

EXIT chart



An example EXIT chart showing two components "right" and "left" and an example decoding (blue)

An **Extrinsic information transfer chart**, commonly called an **EXIT chart**, is a technique to aid the construction of good iteratively-decoded error-correcting codes (in particular low-density parity-check (LDPC) codes and Turbo codes).

EXIT charts were developed by Stephan ten Brink, building on the concept of extrinsic information developed in the Turbo coding community. An EXIT chart includes the response of elements of decoder (for example a convolutional decoder of a Turbo code, the LDPC parity-check nodes or the LDPC variable nodes). The response can either be seen as extrinsic information or a representation of the messages in belief propagation.

If there are two components which exchange messages, the behaviour of the decoder can be plotted on a two-dimensional chart. One component is plotted with its input on the horizontal axis and its output on the vertical axis. The other component is plotted with its input on the vertical axis and its output on the horizontal axis. The decoding path followed is found by stepping between the two curves. For a successful decoding, there

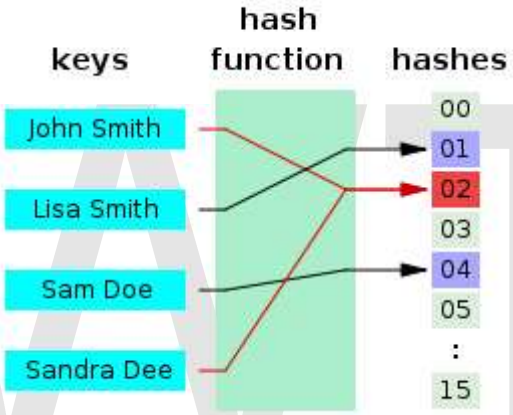
must be a clear swath between the curves so that iterative decoding can proceed from 0 bits of extrinsic information to 1 bit of extrinsic information.

A key assumption is that the messages to and from an element of the decoder can be described by a single number, the extrinsic information. This is true when decoding codes from a binary erasure channel but otherwise the messages are often samples from a Gaussian distribution with the correct extrinsic information. The other key assumption is that the messages are independent (equivalent to an infinite block-size code without local structure between the components)

To make an optimal code, the two transfer curves need to lie close to each other. This observation is supported by the theoretical result that for capacity to be reached for a code over a binary-erasure channel there must be no area between the curves and also by the insight that a large number of iterations are required for information to be spread throughout all bits of a code.

Chapter-15

Hash Function



A hash function that maps names to integers from 0 to 15.. There is a collision between keys "John Smith" and "Sandra Dee".

A **hash function** is any well-defined procedure or mathematical function that converts a large, possibly variable-sized amount of data into a small datum, usually a single integer that may serve as an index to an array (cf. associative array). The values returned by a hash function are called **hash values**, **hash codes**, **hash sums**, **checksums** or simply **hashes**.

Hash functions are mostly used to speed up table lookup or data comparison tasks—such as finding items in a database, detecting duplicated or similar records in a large file, finding similar stretches in DNA sequences, and so on.

A hash function may map two or more keys to the same hash value. In many applications, it is desirable to minimize the occurrence of such collisions, which means that the hash function must map the keys to the hash values as evenly as possible. Depending on the application, other properties may be required as well. Although the idea was conceived in the 1950s, the design of good hash functions is still a topic of active research.

Hash functions are related to (and often confused with) checksums, check digits, fingerprints, randomization functions, error correcting codes, and cryptographic hash functions. Although these concepts overlap to some extent, each has its own uses and requirements and is designed and optimised differently. The HashKeeper database maintained by the American National Drug Intelligence Center, for instance, is more aptly described as a catalog of file fingerprints than of hash values.

Applications

Hash tables

Hash functions are primarily used in hash tables, to quickly locate a data record (for example, a dictionary definition) given its search key (the headword). Specifically, the hash function is used to map the search key to the hash. The index gives the place where the corresponding record should be stored. Hash tables, in turn, are used to implement associative arrays and dynamic sets.

In general, a hashing function may map several different keys to the same index. Therefore, each slot of a hash table is associated with (implicitly or explicitly) a set of records, rather than a single record. For this reason, each slot of a hash table is often called a *bucket*, and hash values are also called *bucket indices*.

Thus, the hash function only hints at the record's location—it tells where one should start looking for it. Still, in a half-full table, a good hash function will typically narrow the search down to only one or two entries.

Caches

Hash functions are also used to build caches for large data sets stored in slow media. A cache is generally simpler than a hashed search table, since any collision can be resolved by discarding or writing back the older of the two colliding items.

Bloom filters

Hash functions are an essential ingredient of the Bloom filter, a compact data structure that provides an enclosing approximation to a set of keys.

Finding duplicate records

When storing records in a large unsorted file, one may use a hash function to map each record to an index into a table T, and collect in each bucket $T[i]$ a list of the numbers of all records with the same hash value i. Once the table is complete, any two duplicate records will end up in the same bucket. The duplicates can then be found by scanning every bucket $T[i]$ which contains two or more members, fetching those records, and comparing them. With a table of appropriate size, this method is likely to be much faster

than any alternative approach (such as sorting the file and comparing all consecutive pairs).

Finding similar records

Hash functions can also be used to locate table records whose key is similar, but not identical, to a given key; or pairs of records in a large file which have similar keys. For that purpose, one needs a hash function that maps similar keys to hash values that differ by at most *m*, where *m* is a small integer (say, 1 or 2). If one builds a table of *T* of all record numbers, using such a hash function, then similar records will end up in the same bucket, or in nearby buckets. Then one need only check the records in each bucket *T*[*i*] against those in buckets *T*[*i*+*k*] where *k* ranges between -*m* and *m*.

This class includes the so-called acoustic fingerprint algorithms, that are used to locate similar-sounding entries in large collection of audio files (as in the MusicBrainz song labeling service). For this application, the hash function must be as insensitive as possible to data capture or transmission errors, and to "trivial" changes such as timing and volume changes, compression, etc.

Finding similar substrings

The same techniques can be used to find equal or similar stretches in a large collection of strings, such as a document repository or a genomic database. In this case, the input strings are broken into many small pieces, and a hash function is used to detect potentially equal pieces, as above.

The Rabin-Karp algorithm is a relatively fast string searching algorithm that works in O(*n*) time on average. It is based on the use of hashing to compare strings.

Geometric hashing

This principle is widely used in computer graphics, computational geometry and many other disciplines, to solve many proximity problems in the plane or in three-dimensional space, such as finding closest pairs in a set of points, similar shapes in a list of shapes, similar images in an image database, and so on. In these applications, the set of all inputs is some sort of metric space, and the hashing function can be interpreted as a partition of that space into a grid of *cells*. The table is often an array with two or more indices (called a *grid file*, *grid index*, *bucket grid*, and similar names), and the hash function returns an index tuple. This special case of hashing is known as geometric hashing or *the grid method*. Geometric hashing is also used in telecommunications (usually under the name vector quantization) to encode and compress multi-dimensional signals.

Properties

Good hash functions, in the original sense of the term, are usually required to satisfy certain properties listed below. Note that different requirements apply to the other related concepts (cryptographic hash functions, checksums, etc.).

Low cost

The cost of computing a hash function must be small enough to make a hashing-based solution more efficient than alternative approaches. For instance, a self-balancing binary tree can locate an item in a sorted table of n items with O(log n) key comparisons. Therefore, a hash table solution will be more efficient than a self-balancing binary tree if the number of items is large and the hash function produces few collisions and less efficient if the number of items is small and the hash function is complex.

Determinism

A hash procedure must be deterministic—meaning that for a given input value it must always generate the same hash value. In other words, it must be a function of the hashed data, in the mathematical sense of the term. This requirement excludes hash functions that depend on external variable parameters, such as pseudo-random number generators or the time of day. It also excludes functions that depend on the memory address of the object being hashed, because that address may change during execution (as may happen on systems that use certain methods of garbage collection), although sometimes rehashing of the item is possible).

Uniformity

A good hash function should map the expected inputs as evenly as possible over its output range. That is, every hash value in the output range should be generated with roughly the same probability. The reason for this last requirement is that the cost of hashing-based methods goes up sharply as the number of *collisions*—pairs of inputs that are mapped to the same hash value—increases. Basically, if some hash values are more likely to occur than others, a larger fraction of the lookup operations will have to search through a larger set of colliding table entries.

Note that this criterion only requires the value to be *uniformly distributed*, not *random* in any sense. A good randomizing function is (barring computational efficiency concerns) generally a good choice as a hash function, but the converse need not be true.

Hash tables often contain only a small subset of the valid inputs. For instance, a club membership list may contain only a hundred or so member names, out of the very large set of all possible names. In these cases, the uniformity criterion should hold for almost all typical subsets of entries that may be found in the table, not just for the global set of all possible entries.

In other words, if a typical set of *m* records is hashed to *n* table slots, the probability of a bucket receiving many more than *m/n* records should be vanishingly small. In particular, if *m* is less than *n*, very few buckets should have more than one or two records. (In an ideal "perfect hash function", no bucket should have more than one record; but a small number of collisions is virtually inevitable, even if *n* is much larger than *m*).

When testing a hash function, the uniformity of the distribution of hash values can be evaluated by the chi-square test.

Variable range

In many applications, the range of hash values may be different for each run of the program, or may change along the same run (for instance, when a hash table needs to be expanded). In those situations, one needs a hash function which takes two parameters—the input data *z*, and the number *n* of allowed hash values.

A common solution is to compute a fixed hash function with a very large range (say, 0 to $2^{32}-1$), divide the result by *n*, and use the division's remainder. If *n* is itself a power of 2, this can be done by bit masking and bit shifting. When this approach is used, the hash function must be chosen so that the result has fairly uniform distribution between 0 and $n-1$, for any *n* that may occur in the application. Depending on the function, the remainder may be uniform only for certain *n*, e.g. odd or prime numbers.

It is possible to relax the restriction of the table size being a power of 2 and not having to perform any modulo, remainder or division operation -as these operation are considered computational costly in some contexts. For example, when *n* is significantly less than 2^b begin with a pseudo random number generator (PRNG) function *P*(*key*), uniform on the interval $[0, 2^b-1]$. Consider the ratio $q = 2^b / n$. Now the hash function can be seen as the value of *P*(*key*) / *q*. Rearranging the calculation and replacing the 2^b-division by bit shifting right (>>) *b* times you end up with hash function $n * P(key) >> b$.

Variable range with minimal movement (dynamic hash function)

When the hash function is used to store values in a hash table that outlives the run of the program, and the hash table needs to be expanded or shrunk, the hash table is referred to as a dynamic hash table.

A hash function that will relocate the minimum number of records when the table is resized is desirable. What is needed is a hash function *H(z,n)* – where *z* is the key being hashed and *n* is the number of allowed hash values – such that $H(z,n+1) = H(z,n)$ with probability close to $n/(n+1)$.

Linear hashing and spiral storage are examples of dynamic hash functions that execute in constant time but relax the property of uniformity to achieve the minimal movement property.

Extendible hashing uses a dynamic hash function that requires space proportional to n to compute the hash function, and it becomes a function of the previous keys that have been inserted.

Several algorithms that preserve the uniformity property but require time proportional to n to compute the value of $H(z,n)$ have been invented.

Data normalization

In some applications, the input data may contain features that are irrelevant for comparison purposes. For example, when looking up a personal name, it may be desirable to ignore the distinction between upper and lower case letters. For such data, one must use a hash function that is compatible with the data equivalence criterion being used: that is, any two inputs that are considered equivalent must yield the same hash value. This can be accomplished by normalizing the input before hashing it, as by upper-casing all letters.

Continuity

A hash function that is used to search for similar (as opposed to equivalent) data must be as continuous as possible; two inputs that differ by a little should be mapped to equal or nearly equal hash values.

Note that continuity is usually considered a fatal flaw for checksums, cryptographic hash functions, and other related concepts. Continuity is desirable for hash functions only in some applications, such as hash tables that use linear search.

Hash function algorithms

For most types of hashing functions the choice of the function depends strongly on the nature of the input data, and their probability distribution in the intended application.

Trivial hash function

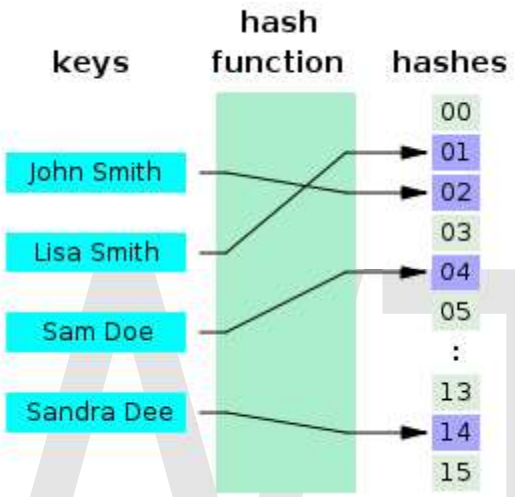
If the datum to be hashed is small enough, one can use the datum itself (reinterpreted as an integer in binary notation) as the hashed value. The cost of computing this "trivial" (identity) hash function is effectively zero.

The meaning of "small enough" depends on how much memory is available for the hash table. A typical PC (as of 2008) might have a gigabyte of available memory, meaning that hash values of up to 30 bits could be accommodated. However, there are many applications that can get by with much less. For example, when mapping character strings between upper and lower case, one can use the binary encoding of each character, interpreted as an integer, to index a table that gives the alternative form of that character ("A" for "a", "8" for "8", etc.). If each character is stored in 8 bits (as in ASCII or ISO

Latin 1), the table has only $2^8 = 256$ entries; in the case of Unicode characters, the table would have $17 \times 2^{16} = 1114112$ entries.

The same technique can be used to map two-letter country codes like "us" or "za" to country names (26^2=676 table entries), 5-digit zip codes like 13083 to city names (100000 entries), etc. Invalid data values (such as the country code "xx" or the zip code 00000) may be left undefined in the table, or mapped to some appropriate "null" value.

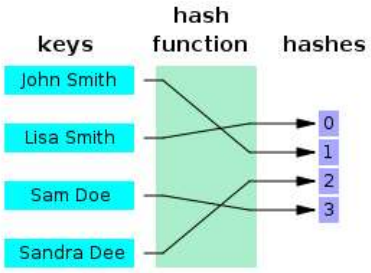
Perfect hashing



A perfect hash function for the four names shown

A hash function that is injective—that is, maps each valid input to a different hash value—is said to be **perfect**. With such a function one can directly locate the desired entry in a hash table, without any additional searching.

Minimal perfect hashing



A minimal perfect hash function for the four names shown

A perfect hash function for n keys is said to be **minimal** if its range consists of n *consecutive* integers, usually from 0 to $n-1$. Besides providing single-step lookup, a minimal perfect hash function also yields a compact hash table, without any vacant slots.

Minimal perfect hash functions are much harder to find than perfect ones with a wider range.

Hashing uniformly distributed data

If the inputs are bounded-length strings (such as telephone numbers, car license plates, invoice numbers, etc.), and each input may independently occur with uniform probability, then a hash function need only map roughly the same number of inputs to each hash value. For instance, suppose that each input is an integer z in the range 0 to $N-1$, and the output must be an integer h in the range 0 to $n-1$, where N is much larger than n. Then the hash function could be $h = z$ **mod** n (the remainder of z divided by n), or $h = (z \times n) \div N$ (the value z scaled down by n/N and truncated to an integer), or many other formulas.

Warning: $h = z$ **mod** n was used in many of the original random number generators, but was found to have a number of issues. One of which is that as n approaches N, this function becomes less and less uniform.

Hashing data with other distributions

These simple formulas will not do if the input values are not equally likely, or are not independent. For instance, most patrons of a supermarket will live in the same geographic area, so their telephone numbers are likely to begin with the same 3 to 4 digits. In that case, if n is 10000 or so, the division formula $(z \times n) \div N$, which depends mainly on the leading digits, will generate a lot of collisions; whereas the remainder formula z **mod** n, which is quite sensitive to the trailing digits, may still yield a fairly even distribution.

Hashing variable-length data

When the data values are long (or variable-length) character strings—such as personal names, web page addresses, or mail messages—their distribution is usually very uneven, with complicated dependencies. For example, text in any natural language has highly non-uniform distributions of characters, and character pairs, very characteristic of the language. For such data, it is prudent to use a hash function that depends on all characters of the string—and depends on each character in a different way.

In cryptographic hash functions, a Merkle–Damgård construction is usually used. In general, the scheme for hashing such data is to break the input into a sequence of small units (bits, bytes, words, etc.) and combine all the units b, b, ..., $b[m]$ sequentially, as follows

```
S ← S0;                          // Initialize the state.
for k in 1, 2, ..., m do          // Scan the input data units:
  S ← F(S, b[k]);                //   Combine data unit k into the
state.
return G(S, n)                   // Extract the hash value from the
state.
```

This schema is also used in many text checksum and fingerprint algorithms. The state variable *S* may be a 32- or 64-bit unsigned integer; in that case, *S0* can be 0, and *G*(*S*,*n*) can be just *S* **mod** *n*. The best choice of *F* is a complex issue and depends on the nature of the data. If the units *b*[*k*] are single bits, then *F*(*S*,*b*) could be, for instance

```
if highbit(S) = 0 then
  return 2 * S + b
else
  return (2 * S + b) ^ P
```

Here *highbit*(*S*) denotes the most significant bit of *S*; the '*' operator denotes unsigned integer multiplication with lost overflow; '^' is the bitwise exclusive or operation applied to words; and *P* is a suitable fixed word.

Special-purpose hash functions

In many cases, one can design a special-purpose (heuristic) hash function that yields many fewer collisions than a good general-purpose hash function. For example, suppose that the input data are file names such as `FILE0000.CHK`, `FILE0001.CHK`, `FILE0002.CHK`, etc., with mostly sequential numbers. For such data, a function that extracts the numeric part *k* of the file name and returns *k* **mod** *n* would be nearly optimal. Needless to say, a function that is exceptionally good for a specific kind of data may have dismal performance on data with different distribution.

Rolling hash

In some applications, such as substring search, one must compute a hash function *h* for every *k*-character substring of a given *n*-character string *t*; where *k* is a fixed integer, and *n* is *k*. The straightforward solution, which is to extract every such substring *s* of *t* and compute *h*(*s*) separately, requires a number of operations proportional to *k*·*n*. However, with the proper choice of *h*, one can use the technique of rolling hash to compute all those hashes with an effort proportional to *k*+*n*.

Universal hashing

A universal hashing scheme is a randomized algorithm that selects a hashing function *h* among a family of such functions, in such a way that the probability of a collision of any two distinct keys is 1/*n*, where *n* is the number of distinct hash values desired—independently of the two keys. Universal hashing ensures (in a probabilistic sense) that the hash function application will behave as well as if it were using a random function, for any distribution of the input data. It will however have more collisions than perfect hashing, and may require more operations than a special-purpose hash function.

Hashing with checksum functions

One can adapt certain checksum or fingerprinting algorithms for use as hash functions. Some of those algorithms will map arbitrary long string data *z*, with any typical real-

world distribution—no matter how non-uniform and dependent—to a 32-bit or 64-bit string, from which one can extract a hash value in 0 through $n-1$.

This method may produce a sufficiently uniform distribution of hash values, as long as the hash range size n is small compared to the range of the checksum or fingerprint function. However, some checksums fare poorly in the avalanche test, which may be a concern in some applications. In particular, the popular CRC32 checksum provides only 16 bits (the higher half of the result) that are usable for hashing. Moreover, each bit of the input has a deterministic effect on each bit of the CRC32, that is one can tell without looking at the rest of the input, which bits of the output will flip if the input bit is flipped; so care must be taken to use all 32 bits when computing the hash from the checksum.

Hashing with cryptographic hash functions

Some cryptographic hash functions, such as SHA-1, have even stronger uniformity guarantees than checksums or fingerprints, and thus can provide very good general-purpose hashing functions.

In ordinary applications, this advantage may be too small to offset their much higher cost. However, this method can provide uniformly distributed hashes even when the keys are chosen by a malicious agent. This feature may help protect services against denial of service attacks.

Origins of the term

The term "hash" comes by way of analogy with its non-technical meaning, to "chop and mix". Indeed, typical hash functions, like the **mod** operation, "chop" the input domain into many sub-domains that get "mixed" into the output range to improve the uniformity of the key distribution.

Donald Knuth notes that Hans Peter Luhn of IBM appears to have been the first to use the concept, in a memo dated January 1953, and that Robert Morris used the term in a survey paper in CACM which elevated the term from technical jargon to formal terminology.

List of hash functions

- Bernstein hash
- Fowler-Noll-Vo hash function (32, 64, 128, 256, 512, or 1024 bits)
- Jenkins hash function (32 bits)
- Pearson hashing (8 bits)
- Zobrist hashing

Chapter-16

Group Code Recording and Binary Golay Code

Group code recording

In computer science, **group code recording (GCR)** refers to several distinct but related encoding methods for magnetic media. The first, used in 6250 cpi magnetic tape, is an error-correcting code combined with a run length limited encoding scheme. The others are different floppy disk encoding methods used in some microcomputers until the late 1980s.

GCR for 9-track reel-to-reel tape

In order to reliably read and write to magnetic tape, several constraints on the signal to be written must be followed. The first is that two adjacent flux reversals must be separated by a certain distance on the media. The second is that there must be a flux reversal often enough to keep the reader's clock in phase with the written signal; that is, the signal must be self-clocking. Prior to 6250 cpi tapes, 1600 cpi tapes satisfied these constraints using a technique called phase encoding, which was only 50% efficient. For 6250 GCR tapes, a (0,2)RLL code is used. This code requires five bits to be written for every four bits of data. The code is structured so that no more than two zero bits (which are represented by lack of a flux reversal) can occur in a row, either within a code or between codes, no matter what the data was. This RLL code is applied independently to the data going to each of the 9 tracks.

Of the 32 5-bit patterns, 8 begin with two consecutive zero bits, 6 others end with two consecutive zero bits, and one more (10001) contains three consecutive zero bits. Removing the all-ones pattern (11111) from the remainder leaves 16 suitable code words.

The 6250 GCR RLL code:

Nibble	Code	Nibble	Code
0000	11001	1000	11010
0001	11011	1001	01001
0010	10010	1010	01010

0011	10011	1011	01011
0100	11101	1100	11110
0101	10101	1101	01101
0110	10110	1110	01110
0111	10111	1111	01111

11 of the nibbles (other than xx00 and 0001) have their code formed by prepending the complement of the msbit. The other 5 values are assigned codes beginning with 11. Nibbles of the form ab00 have codes 11baa□, i.e. the bit reverse of the code for ab11. The code 0001 is assigned the remaining value 11011.

Because of the extremely high density of 6250 cpi tape, the RLL code is not sufficient to ensure reliable data storage. On top of the RLL code, an error-correcting code called the Optimal Rectangular Code (ORC) is applied. This code is a combination of a parity track and polynomial code similar to a CRC, but structured for error correction rather than error detection. For every 7 bytes written to the tape (before RLL encoding), an 8th check byte is calculated and written to the tape. When reading, the parity is calculated on each byte and exclusive-or'd with the contents of the parity track, and the polynomial check code calculated and exclusive-or'd with the received check code, resulting in two 8-bit syndrome words. If these are both zero, the data is error free. Otherwise, error-correction logic in the tape controller corrects the data before it is forwarded to the host. The error correcting code is able to correct any number of errors in any single track, or in any two tracks if the erroneous tracks can be identified by other means.

IBM documents refer to the error correcting code itself as "group coded recording". However, GCR has come to refer to the recording format of 6250 cpi tape as a whole, and later to formats which use similar RLL codes without the error correction code.

GCR for floppy disks

Like magnetic tape drives, floppy disk drives have physical limits on the spacing of flux reversals (also called transitions, represented by 1 bits).

For the Apple II floppy drive, Steve Wozniak invented a floppy controller which (along with the drive itself) imposed two constraints

- Between any two one bits, there may be a maximum of one zero bit.
- Each 8-bit byte must start with a one bit.

The simplest scheme to ensure compliance with these limits is to record an extra "clock" transition between each data bit. This scheme is called FM (Frequency Modulation) or "4 and 4", and allows only 10 256-byte sectors per track to be recorded on a single-density 5¼ floppy.

Wozniak realized that a more complex encoding scheme would allow each 8-bit byte on disk to hold 5 bits of useful data rather than 4 bits. This is because there are 34 bytes which have the top bit set and no two zero bits in a row. This encoding scheme became known as "5 and 3" encoding, and allowed 13 sectors per track; it was used for Apple DOS 3.1, 3.2, and 3.2.1, as well as for the earliest version of Apple CP/M. Later, the design of the floppy drive controller was modified to allow a byte on disk to contain exactly one pair of zero bits in a row. This allowed each 8-bit byte to hold 6 bits of useful data, and allowed 16 sectors per track. This scheme is known as "6 and 2", and was used on Apple Pascal, Apple DOS 3.3 and ProDOS, and later on the 400K and 800K 3½ disks on the Macintosh and Apple II. Apple did not originally call this scheme "GCR", but the term was later applied to it to distinguish it from IBM PC floppies which used the MFM encoding scheme.

Independently, Commodore Business Machines created a Group Code Recording scheme for their Commodore 2040 floppy disk drive (launched in the spring of 1979). The relevant constraint on the 2040 drive was that no more than two zero bits could occur in a row, with no special constraint on the first bit in a byte. This allowed the use of a scheme similar to that used in 6250 tape drives. Every 4 bits of data are translated into 5 bits on disk, according to the following table:

Nibble	Code	Nibble	Code
0000	01010	1000	01001
0001	01011	1001	11001
0010	10010	1010	11010
0011	10011	1011	11011
0100	01110	1100	01101
0101	01111	1101	11101
0110	10110	1110	11110
0111	10111	1111	10101

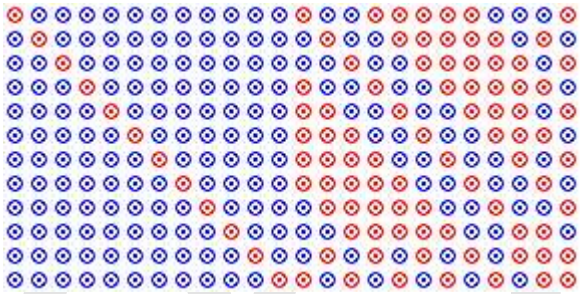
Note no code starts with two zero bits, nor ends with two zero bits. This ensures that regardless of the input data, the encoded data will never contain more than two zero bits in a row. Also note that with this encoding not more than eight one bits in a row are possible. Therefore Commodore used sequences of ten or more one bits in a row as synchronization mark.

Partially because of this more efficient scheme, Commodore was able to fit 170KB on a standard single-density floppy, where Apple fit 140K (6 and 2) or 114K (5 and 3) and an FM-encoded floppy held only 88K.

Binary Golay code

In mathematics and electronics engineering, a **binary Golay code** is a type of error-correcting code used in digital communications. The binary Golay code, along with the ternary Golay code, has a particularly deep and interesting connection to the theory of finite sporadic groups in mathematics. These codes are named in honor of Marcel J. E. Golay.

There are two closely-related binary Golay codes. The **extended binary Golay code** encodes 12 bits of data in a 24-bit word in such a way that any triple-bit error can be corrected and any quadruple-bit error can be detected. The other, the **perfect binary Golay code**, has codewords of length 23 and is obtained from the extended binary Golay code by deleting one coordinate position (conversely, the extended binary Golay code is obtained from the perfect binary Golay code by adding a parity bit). In standard code notation the codes have parameters [24, 12, 8] and [23, 12, 7].



Generator matrix for the extended binary Golay code

Mathematical definition

In mathematical terms, the extended binary Golay code consists of a 12-dimensional subspace W of the space $V=\mathbf{F}_2^{24}$ of 24-bit words such that any two distinct elements of W differ in at least eight coordinates. Equivalently, any non-zero element of W has at least eight non-zero coordinates.

- The possible sets of non-zero coordinates as w ranges over W are called *code words*. In the extended binary Golay code, all code words have the Hamming weights of 0, 8, 12, 16, or 24.
- Up to relabeling coordinates, W is unique.

The perfect binary Golay code is a perfect code. That is, the spheres of radius three around code words form a partition of the vector space.

The automorphism group of the binary Golay code is the Mathieu group M_{23}. The automorphism group of the extended binary Golay code is the Mathieu group M_{24}. The other Mathieu groups occur as stabilizers of one or several elements of W.

The Golay code words of weight eight are elements of the S(5,8,24) Steiner system.

Constructions

1. Lexicographic code: Order the vectors in V lexicographically (i.e., interpret them as unsigned 24-bit binary integers and take the usual ordering). Starting with $w_1 = 0$, define w_2, w_3, ..., w_{12} by the rule that w_n is the smallest integer which differs from all linear combinations of previous elements in at least eight coordinates. Then W can be defined as the span of w_1, ..., w_{12}.
2. Quadratic residue code: Consider the set N of quadratic non-residues (mod 23). This is an 11-element subset of the cyclic group $\mathbf{Z}/23\mathbf{Z}$. Consider the translates $t+N$ of this subset. Augment each translate to a 12-element set S_t by adding an element ∞. Then labeling the basis elements of V by 0, 1, 2, ..., 22, ∞, W can be defined as the span of the words S_t together with the word consisting of all basis vectors. (The perfect code is obtained by leaving out ∞.)
3. As a Cyclic code: The perfect G_{23} code can be constructed via factorization of $x^{23} - 1$, it is the code generated by $x^{11} + x^{10} + x^6 + x^5 + x^4 + x^2 + 1 \mid x^{23} - 1$
4. The Miracle Octad Generator of R. T. Curtis: This uses a 4×6 array of square cells to picture the 759 Hamming-weight-8 code words, or "octads," of the extended binary Golay code. The remaining code words are obtained via symmetric differences of subsets of the 24 cells-- i.e., by binary addition.
5. Winning positions in the mathematical game of Mogul: a position in Mogul is a row of 24 coins. Each turn consists of flipping from one to seven coins such that the leftmost of the flipped coins goes from head to tail. The losing positions are those with no legal move. If heads are interpreted as 1 and tails as 0 then moving to a codeword from the extended binary Golay code guarantees it will be possible to force a win.
6. A generator matrix for the binary Golay code is **I A**, where **I** is the 12×12 identity matrix, and **A** is the complement of the adjacency matrix of the icosahedron.

Practical applications of Golay codes

NASA Deep Space Missions

The Voyager 1 & 2 spacecraft needed to transmit hundreds of color pictures of Jupiter and Saturn in their 1979, 1980, and 1981 fly-bys within a constrained telecommunications bandwidth.

- Color image transmission required three times the amount of data, so the Golay (24,12,8) code was used.
- This Golay code is only triple-error correcting, but it could be transmitted at a much higher data rate than the Hadamard code that was used during the Mariner mission.

ALE HF data communications

The new American government standards for automatic link establishment (ALE) in High Frequency (HF) radio systems specifies the use of an extended (24,12) Golay block code for forward error correction (FEC).

- The Extended (24,12) Golay Code specified is a (24,12) block code.
- This code encodes 12 data bits to produce 24-bit code words.
- It is furthermore a systematic code, meaning that the 12 data bits are present in unchanged form in the code word.

The minimum Hamming distance between any two code words (the number of bits by which any pair of code words differs) is eight.

Chapter-17

Casting out Nines and Echo (computing)

Casting out nines

Casting out nines is a sanity check to ensure that hand computations of sums, differences, products, and quotients of integers are correct. By looking at the digital roots of the inputs and outputs, the casting-out-nines method can help one check arithmetic calculations. The method is so simple that most schoolchildren can apply it without understanding its mathematical underpinnings.

Examples

The method involves converting each number into its "casting-out-nines" equivalent, and then redoing the arithmetic. The casting-out-nines answer should equal the casting-out-nines version of the original answer. Below are examples for using casting out nines to check addition, subtraction, multiplication, and division.

Addition

In each addend, cross out all 9's and pairs of digits that total 9, then add together what remains. These new values are called excesses. Add up leftover digits for each addend until one digit is reached. Now process the sum and also the excesses to get a *final* excess.

$3264 \Rightarrow 6$ 2 and 4 add up to 6

$8415 \Rightarrow 0$ There are no digits left.

$2946 \Rightarrow 3$ 2, 4, and 6 make 12; 1 and 2 make 3.

$+3206 \Rightarrow 2$ 2 and 0 are 2.

$17831 \Rightarrow \Vert$ 7, 3, and 1 make 11; 1 and 1 add up to 2.

$\Downarrow \quad \Downarrow$

$2 \quad \Leftrightarrow 2$ The excess from the sum should equal the final excess from the addends.

Subtraction

$5643 \Rightarrow 0(9)$ First, cross out all 9's and digits that total 9 in both minuend and subtrahend (italicized).

$\underline{-2891} \Rightarrow -2$ Add up leftover digits for each value until one digit is reached.

2752 Now follow the same procedure with the difference, coming to a single digit.

$\Downarrow \quad \Downarrow$ Because subtracting 2 from zero gives a negative number, borrow a 9 from the minuend.

$7 \quad \Leftrightarrow 7$ The difference between the minuend and the subtrahend excesses should equal the difference excess.

Multiplication

$548 \Rightarrow 8$ First, cross out all 9's and digits that total 9 in each factor (italicized).

$\underline{\times 629} \Rightarrow 8$ Add up leftover digits for each multiplicand until one digit is reached.

344692 Multiply the two excesses, and then add until one digit is reached.

$\Downarrow \quad \Downarrow$ Do the same with the product, crossing out 9's and getting one digit.

$1 \quad \Leftrightarrow 1^{*}$ The excess from the product should equal the final excess from the factors.

*8 times 8 is 64; 6 and 4 are 10; 1 and 0 are 1

Division

$275462 \div 877 = 314\ r.\ 84$ Cross out all 9's and digits that total 9 in the divisor, quotient, and remainder.

$\Downarrow \quad \Downarrow \quad \Downarrow \quad \Downarrow$ Add up all uncrossed digits from each value to a single digits.

$8 \quad \Leftrightarrow \quad 4 \times 8 + 3$ The dividend excess should equal the final excess from the other values.

How it works

The method works because the original numbers are 'decimal' (base 10), the modulus is chosen to differ by 1, and casting out is equivalent to taking a digit sum. In general any two 'large' integers, x and y, expressed in any smaller *modulus* as x' and y' (for example, modulo 7) will always have the same sum, difference, product or quotient as their originals. But this property is also preserved for the 'digit sum' where the 'base' and the 'modulus' differ by 1.

To see this take an example: Both 900 and 630 are exactly divisible by 9 and have the same digit sum - '63' changes into '90' by repeated addition of '09' and the change in the second digit always offsets the change in the first ('63' to '72' to '81' to '90'). For two 'decimal' numbers not generally congruent modulo 9 (like '914' and '673') the picture remains the same; we can consider their congruences *pairwise* ('900' with '630') plus ('09' with '36') plus ('5' with '7'). Thus the digit sum of any number and its congruence modulo 9 are always fixed in base 10.

If a calculation was correct before casting out, casting out on both sides will preserve correctness. However, it is possible that two previously unequal integers will be identical modulo 9 (on average, a ninth of the time).

One should note that the operation does not work on fractions, since a given fractional number does not have a unique representation.

A variation on the explanation

A nice trick for very young children to learn to add nine is to add ten to the digit and to count back one. Since we are adding 1 to the ten's digit and subtracting one from the unit's digit, the sum of the digits should remain the same. For example 9+2=11 with 1+1=2. When adding 9 to itself, we would thus expect the sum of the digits to be 9 as follows: 9+9=18 (1+8=9) and 9+9+9=27 (2+7=9). Let us look at a simple multiplication: 5×7=35 (3+5=8). Now consider (7+9)×5=16×5=80 (8+0=8) or 7×(9+5)=7×14=98 (9+8=17 1+7=8).

Any positive integer can be written as $9 \times n + a$ where 'a' is a single digit 0 to 8 and 'n' is any positive integer. Thus, using the distributive rule $(9 \times n + a)\times(9 \times m + b)= 9 \times 9 \times n \times m + 9 \times(am+bn) +ab$. Since the first two factors are multiplied by 9, their sums will end up being 9 or 0, leaving us with 'ab'. In our example, 'a' was 7 and 'b' was 5. We would expect in any base system the number before that base would behave just like the nine.

History

Abjectio novenaria (Latin for "casting out nines") was known to the Roman bishop Hippolytos as early as the third century. It was employed by Twelfth-century Hindu mathematicians. in the 17th century, Gottfried Wilhelm Leibniz not only used the method extensively, but presented it frequently as a model for rationality: "By means of this, once a reasoning in morality, physics, medicine or metaphysics is reduced to these terms or characters, one will be able to apply to it at any moment a numerical test, so that it will be impossible to be mistaken if one does not so desire...".

Synergetics, R. Buckminster Fuller claims to have used casting out nines "before World War I." Fuller explains how to cast out nines and makes other claims about the resulting 'indigs,' but he fails to note that casting out nines can result in false positives.

The method bears striking resemblance to standard signal processing and computational error detection and error correction methods, typically using similar modular arithmetic in checksums and simpler check digits.

Echo

In computer telecommunications, **echo** is the display or return of sent data at or to the sending end of a transmission. Echo can be either **local echo**, where the sending device itself displays the sent data, or **remote echo**, where the receiving device returns the sent data that it receives to the sender (which is of course simply *no* local echo from the point of view of the sending device itself). That latter, when used as a form of error detection to determine that data received at the remote end of a communications line are the same as data sent, is also known as **echoplex**, **echo check**, or **loop check**. When two modems are communicating in **echoplex mode**, for example, the remote modem echoes whatever it receives from the local modem.

Terminological confusion: echo is not duplex

Whilst local echo and remote echo are sometimes referred to as "half-duplex" and "full-duplex", those latter appellations are strictly incorrect, and are misleading. for their correct meanings.) Whilst local echo is *often used with* half-duplex transmission, so that bandwidth is not wasted upon remotely echoing data from the remote end of the communications channel, it is not *the same as* half-duplex transmission. A full-duplex communications channel can still employ local echo, and a half-duplex channel can still (albeit wastefully) employ remote echo, or no echo at all.

Indeed, for example, that is the case for echoplex error checking, which *requires* full-duplex communication, so that received data can be echoed back as they are being received.

Similarly, for another example, in the case of the TELNET communications protocol a local echo protocol operates on top of a full-duplex underlying protocol. The TCP connection over which the TELNET protocol is layered provides a full-duplex connection, with no echo, across which data may be sent in either direction simultaneously. Whereas the *Network Virtual Terminal* that the TELNET protocol itself incorporates is a half-duplex device with (by default) local echo.

The devices that echo locally

terminals are one of the things that may perform echoing for a connection. Others include modems, some form of intervening communications processor, or even the host system itself. For several common computer operating systems, it is the host system itself that performs the echoing, if appropriate (which it isn't for, say, entry of a user password when a terminal first connects and a user is prompted to log in). On OpenVMS, for example, echoing is performed as necessary by the host system. Similarly, on BSD

version 4, local echo is performed by the operating system kernel's *terminal device driver*, according to the state of a device control flag, maintained in software and alterable by applications programs via an `ioctl()` system call. The actual terminals and modems connected to such systems should have *their* local echo facilities switched off (so that they operate in *no echo* mode), lest passwords be locally echoed at password prompts, and all other input appear echoed twice. This is as true for terminal emulator programs, such as C-Kermit, running on a computer as it is for real terminals.

Controlling local echo

Terminal emulators

Most terminal emulator programs have the ability to perform echo locally (which sometimes they mis-name "half-duplex"):

- In the C-Kermit terminal emulator program, local echo is controlled by the `SET TERMINAL ECHO` command, which can be either `SET TERMINAL ECHO LOCAL` (which enables local echoing within the terminal emulator program itself) or `SET TERMINAL ECHO REMOTE` (where disables local echoing, leaving that up to another device in the communications channel — be that the modem or the remote host system — to perform as appropriate).
- In ProComm it is the `Alt`+`E` combination, which is a hot key that may be used at any time to toggle local echo on and off.
- In the Terminal program that came with Microsoft Windows 3.1, local echo is controlled by a checkbox in the "Terminal Preferences" dialogue box accessed from the menu of the terminal program's window.

Modems

In the Hayes command set, the `AT` command that switches local echo off (in command mode) is `E0` and the command to switch it on is `E1`. For local echo in data mode, the commands are `F1` and `F0` respectively. (Note the reversal of the numbers. The "F" commands are not part of the EIA/TIA-602 standard, but the "E" commands are.)

Host systems

Some host systems perform local echo themselves, in their device drivers and so forth.

- In Unix and POSIX-compatible systems, local echo is a flag in the POSIX terminal interface, settable programmatically with the tcsetattr() function.[fn 1] The echoing is performed by the operating system's terminal device (in some way that is not specified by the POSIX standard). The standard utility program that alters this flag programmatically is the stty command, using which the flag may be altered from shell scripts or an interactive shell. The command to turn local echo

(by the host system) on is stty echo and the command to turn it off is stty -echo.[fn 2]

- On OpenVMS systems, the operating system's terminal driver normally performs echoing. The *terminal characteristic* that controls whether it does this is the `ECHO` characteristic, settable with the DCL command SET TERMINAL /ECHO and unsettable with SET TERMINAL /NOECHO.

Chapter-18

BCH Code

In coding theory the **BCH codes** form a class of parameterised error-correcting codes which have been the subject of much academic attention in the last fifty years. BCH codes were invented in 1959 by Hocquenghem, and independently in 1960 by Bose and Ray-Chaudhuri. The acronym *BCH* comprises the initials of these inventors' names.

The principal advantage of BCH codes is the ease with which they can be decoded, via an elegant algebraic method known as syndrome decoding. This allows very simple electronic hardware to perform the task, obviating the need for a computer, and meaning that a decoding device may be made small and low-powered. As a class of codes, they are also highly flexible, allowing control over block length and acceptable error thresholds, meaning that a custom code can be designed to a given specification (subject to mathematical constraints).

Reed–Solomon codes, which are BCH codes, are used in applications such as satellite communications, compact disc players, DVDs, disk drives, and two-dimensional bar codes.

In technical terms a BCH code is a multilevel cyclic variable-length digital error-correcting code used to correct multiple random error patterns. BCH codes may also be used with multilevel phase-shift keying whenever the number of levels is a prime number or a power of a prime number. A BCH code in 11 levels has been used to represent the 10 decimal digits plus a sign digit.

BCH codes are also useful in theoretical computer science, for instance in the MAXEkSAT problem.

Construction

A BCH code is a polynomial code over a finite field with a particularly chosen generator polynomial. It is also a cyclic code.

Simplified BCH codes

For ease of exposition, we first describe a special class of BCH codes. General BCH codes are described in the next section.

Definition. Fix a finite field $GF(q^m)$, where q is a prime. Also fix positive integers n, and d such that $n = q^m - 1$ and $2 \leq d \leq n$. We will construct a polynomial code over $GF(q)$ with code length n, whose minimum Hamming distance is at least d. What remains to be specified is the generator polynomial of this code.

Let α be a primitive nth root of unity in $GF(q^m)$. For all i, let $m_i(x)$ be the minimal polynomial of α^i with coefficients in $GF(q)$. The generator polynomial of the BCH code is defined as the least common multiple $g(x) = \operatorname{lcm}(m_1(x), \ldots, m_{d-1}(x))$.

Example

Let $q = 2$ and $m = 4$ (therefore $n = 15$). We will consider different values of d. There is a primitive root $\alpha \in GF(16)$ satisfying

$$\alpha^4 + \alpha + 1 = 0 \quad (1)$$

its minimal polynomial over $GF(2)$ is :$m_1(x) = x^4 + x + 1$. Note that in $GF(2^4)$, the equation $(a + b)^2 = a^2 + 2ab + b^2 = a^2 + b^2$ holds, and therefore $m_1(\alpha^2) = m_1(\alpha)^2 = 0$. Thus α^2 is a root of $m_1(x)$, and therefore

$$m_2(x) = m_1(x) = x^4 + x + 1.$$

To compute $m_3(x)$, notice that, by repeated application of (1), we have the following linear relations:

- $1 = 0\alpha^3 + 0\alpha^2 + 0\alpha + 1$
- $\alpha^3 = 1\alpha^3 + 0\alpha^2 + 0\alpha + 0$
- $\alpha^6 = 1\alpha^3 + 1\alpha^2 + 0\alpha + 0$
- $\alpha^9 = 1\alpha^3 + 0\alpha^2 + 1\alpha + 0$
- $\alpha^{12} = 1\alpha^3 + 1\alpha^2 + 1\alpha + 1$

Five right-hand-sides of length four must be linearly dependent, and indeed we find a linear dependency $\alpha^{12} + \alpha^9 + \alpha^6 + \alpha^3 + 1 = 0$. Since there is no smaller degree dependency, the minimal polynomial of α^3 is :$m_3(x) = x^4 + x^3 + x^2 + x + 1$. Continuing in a similar manner, we find

$$\begin{aligned} m_4(x) &= m_2(x) = m_1(x) = x^4 + x + 1, \\ m_5(x) &= x^2 + x + 1, \\ m_6(x) &= m_3(x) = x^4 + x^3 + x^2 + x + 1, \\ m_7(x) &= x^4 + x^3 + 1. \end{aligned}$$

The BCH code with $d = 1,2,3$ has generator polynomial

$$g(x) = m_1(x) = x^4 + x + 1.$$

It has minimal Hamming distance at least 3 and corrects up to 1 error. Since the generator polynomial is of degree 4, this code has 11 data bits and 4 checksum bits.

The BCH code with $d = 4,5$ has generator polynomial

$$g(x) = \operatorname{lcm}(m_1(x), m_3(x)) = (x^4+x+1)(x^4+x^3+x^2+x+1) = x^8+x^7+x^6+x^4+1.$$

It has minimal Hamming distance at least 5 and corrects up to 2 errors. Since the generator polynomial is of degree 8, this code has 7 data bits and 8 checksum bits.

The BCH code with $d = 6,7$ has generator polynomial

$$\begin{aligned} g(x) &= \operatorname{lcm}(m_1(x), m_3(x), m_5(x)) \\ &= (x^4 + x + 1)(x^4 + x^3 + x^2 + x + 1)(x^2 + x + 1) \\ &= x^{10} + x^8 + x^5 + x^4 + x^2 + x + 1. \end{aligned}$$

It has minimal Hamming distance at least 7 and corrects up to 3 errors. This code has 5 data bits and 10 checksum bits.

The BCH code with $d = 8$ and higher have generator polynomial

$$\begin{aligned} g(x) &= \operatorname{lcm}(m_1(x), m_3(x), m_5(x), m_7(x)) \\ &= (x^4 + x + 1)(x^4 + x^3 + x^2 + x + 1)(x^2 + x + 1)(x^4 + x^3 + 1) \\ &= x^{14} + x^{13} + x^{12} + \cdots + x^2 + x + 1. \end{aligned}$$

This code has minimal Hamming distance 8 and corrects up to 3 errors. It has 1 data bit and 14 checksum bits. In fact, this code has only two codewords: 000000000000000 and 111111111111111.

General BCH codes

General BCH codes differ from the simplified case discussed above in two respects. First, one replaces the requirement $n = q^m - 1$ by a more general condition. Second, the consecutive roots of the generator polynomial may run from $\alpha^c, \ldots, \alpha^{c+d-2}$ instead of $\alpha, \ldots, \alpha^{d-1}$.

Definition. Fix a finite field $GF(q)$, where q is a prime power. Choose positive integers m,n,d,c such that $2 \leq d \leq n$, $\gcd(n,q) = 1$, and m is the multiplicative order of q modulo n.

As before, let α be a primitive nth root of unity in $GF(q^m)$, and let $m_i(x)$ be the minimal polynomial over $GF(q)$ of α^i for all i. The generator polynomial of the BCH code is defined as the least common multiple $g(x) = \operatorname{lcm}(m_c(x), \ldots, m_{c+d-2}(x))$.

Note: if $n = q^m - 1$ as in the simplified definition, then gcd(n,q) is automatically 1, and the order of q modulo n is automatically m. Therefore, the simplified definition is indeed a special case of the general one.

Properties

1. The generator polynomial of a BCH code has degree at most $(d - 1)m$. Moreover, if $q = 2$ and $c = 1$, the generator polynomial has degree at most $dm / 2$.

> Proof: each minimal polynomial $m_i(x)$ has degree at most m. Therefore, the least common multiple of $d - 1$ of them has degree at most $(d - 1)m$. Moreover, if $q = 2$, then $m_i(x) = m_{2i}(x)$ for all i. Therefore, $g(x)$ is the least common multiple of at most $d / 2$ minimal polynomials $m_i(x)$ for odd indices i, each of degree at most m.

2. A BCH code has minimal Hamming distance at least d. Proof: We only give the proof in the simplified case; the general case is similar. Suppose that $p(x)$ is a code word with fewer than d non-zero terms. Then

$$p(x) = b_1 x^{k_1} + \cdots + b_{d-1} x^{k_{d-1}}, \text{ where } k_1 < k_2 < \cdots < k_{d-1}.$$

Recall that $\alpha^1, \ldots, \alpha^{d-1}$ are roots of $g(x)$, hence of $p(x)$. This implies that $b_1, \ldots, b_{d-1}$ satisfy the following equations, for $i = 1, \ldots, d-1$:

$$p(\alpha^i) = b_1 \alpha^{ik_1} + b_2 \alpha^{ik_2} + \cdots + b_{d-1} \alpha^{ik_{d-1}} = 0$$

In matrix form, we have

$$\begin{bmatrix} \alpha^{k_1} & \alpha^{k_2} & \cdots & \alpha^{k_{d-1}} \\ \alpha^{2k_1} & \alpha^{2k_2} & \cdots & \alpha^{2k_{d-1}} \\ \vdots & \vdots & & \vdots \\ \alpha^{(d-1)k_1} & \alpha^{(d-1)k_2} & \cdots & \alpha^{(d-1)k_{d-1}} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_{d-1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

The determinant of this matrix equals

$$\left(\prod_{i=1}^{d-1}\alpha^{k_i}\right)\det\begin{pmatrix}1 & 1 & \cdots & 1\\ \alpha^{k_1} & \alpha^{k_2} & \cdots & \alpha^{k_{d-1}}\\ \vdots & \vdots & & \vdots\\ \alpha^{(d-2)k_1} & \alpha^{(d-2)k_2} & \cdots & \alpha^{(d-2)k_{d-1}}\end{pmatrix} = \left(\prod_{i=1}^{d-1}\alpha^{k_i}\right)\det(V)$$

The matrix V is seen to be a Vandermonde matrix, and its determinant is

$$\det(V) = \prod_{1\le i<j\le d-1}(\alpha^{k_j} - \alpha^{k_i})$$

,

which is non-zero. It therefore follows that $b_1, \ldots, b_{d-1} = 0$, hence $p(x) = 0$.

3. A BCH code is cyclic.

Proof: A polynomial code of length n is cyclic if and only if its generator polynomial divides $x^n - 1$. Since $g(x)$ is the minimal polynomial with roots $\alpha^c, \ldots, \alpha^{c+d-2}$, it suffices to check that each of $\alpha^c, \ldots, \alpha^{c+d-2}$ is a root of $x^n - 1$. This follows immediately from the fact that α is, by definition, an nth root of unity.

Special cases

- A BCH code with $c = 1$ is called a *narrow-sense BCH code*.
- A BCH code with $n = q^m - 1$ is called *primitive*.

Therefore, the "simplified" BCH codes we considered above were just the primitive narrow-sense codes.

- A narrow-sense BCH code with $n = q^m - 1$ is called a *Reed-Solomon code*.

Decoding

There are many algorithms for decoding BCH codes. The most common ones follow this general outline:

1. Calculate the syndrome values for the received vector
2. Calculate the error locator polynomials
3. Calculate the roots of this polynomial to get error location positions.
4. Calculate the error values at these error locations.

Calculate the syndromes

The received vector R is the sum of the correct codeword C and an unknown error vector E. The syndrome values are formed by considering R as a polynomial and evaluating it at $\alpha^c, \ldots, \alpha^{c+d-2}$. Thus the syndromes are

$$s_j = R(\alpha^{c+j-1}) = C(\alpha^{c+j-1}) + E(\alpha^{c+j-1})$$

for $j = 1$ to $d - 1$. Since α^{c+j-1} are the zeros of $g(x)$, of which $C(x)$ is a multiple, $C(\alpha^{c+j-1}) = 0$. Examining the syndrome values thus isolates the error vector so we can begin to solve for it.

If there is no error, $s_j = 0$ for all j. If the syndromes are all zero, then the decoding is done.

Calculate the error location polynomial

If there are nonzero syndromes, then there are errors. The decoder needs to figure out how many errors and the location of those errors.

If there is a single error, write this as $E(x) = e\,x^i$, where i is the location of the error and e is its magnitude. Then the first two syndromes are

$$s_1 = e\,\alpha^{ci}$$
$$s_2 = e\,\alpha^{(c+1)\,i} = \alpha^i s_1$$

so together they allow us to calculate e and provide some information about i (completely determining it in the case of Reed-Solomon codes).

If there are two or more errors,

$$E(x) = e_1 x^{i_1} + e_2 x^{i_2} + \ldots$$

It is not immediately obvious how to begin solving the resulting syndromes for the unknowns e_k and i_k. Two popular algorithms for this task are:

1. Peterson-Gorenstein-Zierler algorithm
2. Berlekamp-Massey algorithm

Peterson Gorenstein Zierler algorithm

Peterson's algorithm is the step 2 of the generalized BCH decoding procedure. We use Peterson's algorithm to calculate the error locator polynomial coefficients $\lambda_1, \lambda_2, \ldots, \lambda_t$ of a polynomial $\Lambda(x) = 1 + \lambda_1 x + \lambda_2 x^2 + \cdots + \lambda_t x^t$

Now the procedure of the Peterson Gorenstein Zierler algorithm for a given (n,k,d_{min}) BCH code designed to correct $[t = \frac{d_{min} - 1}{2}]$ errors is

- First generate the Matrix of $2t$ syndromes
- Next generate the $S_{t \times t}$ matrix with elements that are syndrome values

$$S_{t \times t} = \begin{bmatrix} s_1 & s_2 & s_3 & \dots & s_t \\ s_2 & s_3 & s_4 & \dots & s_{t+1} \\ s_3 & s_4 & s_5 & \dots & s_{t+2} \\ \dots & \dots & \dots & \dots & \dots \\ s_t & s_{t+1} & s_{t+2} & \dots & s_{2t-1} \end{bmatrix}$$

- Generate a c_{tx1} vector with elements

$$C_{t \times 1} = \begin{bmatrix} s_{t+1} \\ s_{t+2} \\ \vdots \\ \vdots \\ s_{2t} \end{bmatrix}$$

- Let Λ denote the unknown polynomial coefficients, which are given by

$$\Lambda_{t \times 1} = \begin{bmatrix} \lambda_t \\ \lambda_{t-1} \\ \vdots \\ \lambda_2 \\ \lambda_1 \end{bmatrix}$$

- Form the matrix equation

$$S_{t \times t} \Lambda_{t \times 1} = C_{t \times 1}$$

- If the determinant of matrix $S_{t \times t}$ exists, then we can actually find an inverse of this matrix and solve for the values of unknown Λ values.
- If $\det(S_{t \times t}) = 0$, then follow

```
if t = 0
```

```
then
      declare an empty error locator polynomial
      stop Peterson procedure.
end
set t ← t − 1
continue from the beginning of Peterson's decoding
```

- After you have values of Λ you have with you the error locator polynomial.
- Stop Peterson procedure.

Factor error locator polynomial

Now that you have the Λ(*x*) polynomial, you can find its roots in the form $\Lambda(x) = (\alpha^i x + 1)(\alpha^j x + 1) \cdots (\alpha^k x + 1)$ using the Chien search algorithm. The exponential powers of the primitive element α will yield the positions where errors occur in the received word; hence the name 'error locator' polynomial.

Calculate error values

Once the error locations are known, the next step is to determine the error values at those locations. The error values are then used to correct the received values at those locations to recover the original codeword.

For the case of binary BCH, this is trivial; just flip the bits for the received word at these positions, and we have the corrected code word. In the more general case, the error weights e_j can be determined by solving the linear system

$$s_1 = e_1 \alpha^{c i_1} + e_2 \alpha^{c i_2} + \ldots$$
$$s_2 = e_1 \alpha^{(c+1) i_1} + e_2 \alpha^{(c+1) i_2} + \ldots$$
$$\ldots$$

However, there is a more efficient method known as the Forney Algorithm, which is based on Lagrange interpolation. First calculate the error evaluator polynomial

$$\omega(x) = S(x)\,\Lambda(x) \pmod{x^{d-1}}$$

Then evaluate the error values.

$$e_j = \frac{\omega(\alpha^{-c i_j})}{\prod_{k \neq j}(1 - \alpha^{c i_k} \alpha^{-c i_j})}$$