# Compression, Encryption and Hashing

## Compression

- Data transmitted across the internet needs to be compressed where possible. Compression reduces file sizes. This:

- Reduce download times.

- Makes best use of bandwidth.

- Reduce file storage requirements.

- Data transmission streams have increased enormously over the years so sending lots of data is becoming less of an issue but expectations are also increasing.

- Reducing the size of data either in storage or in transmission requires various compression techniques, there are many techniques dependent on situation.

- Trade-off of quality of final result for amount of processing power needed to process and decompress.

- Two strategies: lossy and lossless.

**Lossy**

- Lossy compression is a way of reducing file’s size by removing some of the data. The original file cannot be recreated from the compressed file.

- Considerable savings can be made but quality will be affected.

- Typically audio is reduced by a factor of 10:1 and so are photos. Typically videos can be compressed much more 100:1 is common.

- Lossy file formats: JPEG, MPEG and MP3.

- Idea is to remove data that is least important. For example, MP3 removes frequencies humans cannot hear.

- Images will often be reduced in resolution or colour range.

- Video techniques might be:

- Spatial Compression.

- Temporal Compression.

- **Spacial:** takes place on every individual frame of the video, compressing the pixel information as though it was a still image.

- **Temporal:** technique of reducing compressed video size not by encoding each frame as a complete image. Rather frames that are compressed completely are called **key frames** and all other frames are represented by data specifying the change since the last frame called **delta frames.**

- Videos without a lot of motion are best suited to temporal compression.

**Lossless**

- Reduces fie sizes in such a way that no data is lost and the original file can be regenerated exactly.

- Important for programs and files such as spreadsheets.

- Makes use of redundant data so that a data item occurs multiple times the item is stored once along with number of repetitions.

- When storing text each reused word can be stored as a number and then the numbers can be stored as a long text file. The dictionary needs to also be stored with the message.

- Various ways exist to generate dictionaries as a file is parsed but the best-known is the LZW (Lempel-Ziv-Welch) algorithm.

- The dictionary is updated as the file is examined. When a sequence is found that is already in the dictionary, the next character is examined and if this is new, the longer sequence gets added to the dictionary.

- Well known compression formats that use dictionary coding are ZIP, GIF, PNG.

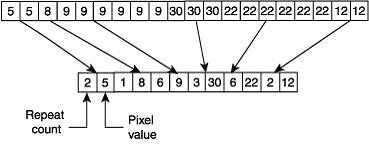
**Run-Length Encoding**

- Another simple approach can be applied to other types of data such as pixels in an image. For example B100 could be 100 blue pixels.

- The image can be reproduced exactly from this information.

- This is known as **Run-Length Encoding (RDE)**

- Easier to implement than Dictionary Coding.



- Lossless compression is rarely as effective as lossy in reducing file size but in some situations that require a faithful reconstruction of the source data it is useful. Ie computer program.

**Summary**

- The usability of web services often requires compression.

- Compression is the reduction of file sizes.

- Lossy compression can achieve big savings but degrades the quality of the data source.

- Lossless compression allows reconstruction of original data source but may increase processing overheads significantly.

## Encryption

- Encryption is the process of scrambling data such that only legitimate users can read it.

- With people spending so much time online they do not want their activities and messages leaked into the public domain.

- Some activities require a higher level of security:

- Online banking.

- Communications involving trade secrets or other sensitive or personal data.

- Encryption changes **plaintext** into **ciphertext.**

- This is done using algorithm and a very large digital number- a **key.**

- two major approaches to encryption: symmetric and asymmetric.

**Symmetric**

- In symmetric, the key used to encrypt a message is also used to decrypt it.

- The key must be long enough to make it very difficult for a computer to brute force hack in a reasonable time.

- Symmetric encryption only works if key is kept secret.

- The rise of the internet makes symmetric key encryption difficult to implement:

- How do you share a key with someone on the other side of the world?

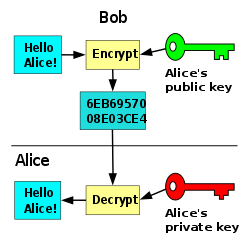
- There is always a danger of a successful attack on symmetric encryption, either by intercepting the key or brute forcing.

**Asymmetric**

- Requires two different keys.

- The key used to encrypt the message is not the same as the key used to decrypt the message.

- This is sometimes known as public/private key system.

-

- the public key can only be used to encrypt messages, but cannot be used to decrypt them.

- It is therefore safe to send it across networks and make it publicly available.

- The private key is never sent across a network and must be kept safe.

- **TLS (Transport Layer Security)** is the most widely used secure protocol on the internet and is used to encrypt data when we see HTTPS.

- Encrypting and decrypting data uses lots of processing time.

- The shorter the key, the less processing time required but the easier it is to hack.

- This is why major companies have been hacked as they try and save money by using a shorter key.

- As computing power expanded, they did not upgrade security.

**Quantum Cryptography**

- In 2007 votes cast in the Swiss canton of Geneva were protected for the first time by quantum cryptography.

0 Both the secret-key and public-key methods have unique flaws. Quantum physics can be used to either solve or expand these flaws.

- Unlike tradition cryptography, quantum cryptography depends on physics not maths.

- Parliamentary elections were transmitted using a secure encryption encoded by a key generated using photons which are packets of light.

- There is little chance it can be cracked using maths as it depends on physics.

- Photos are cool because they can exist in all their possible states at once called the **wave function.**

- This means that whatever direction a photon can **spin** in it does all at once.

- Light in this state is called **unpolarised.**

- LEDs are capable fo creating just one photon at a time and through the use of polarisation filters you can force the photon to take one state or another. **Polarise it.** These are known as Qbits.

- The thing about photons is that once they’re polarised they can’t be accurately measured again except by a filter like the one that produced their current spin.

- So if a photon with a vertical spin is measured through a diagonal filter the photon either won’t pass through or the filter will affect the photon’s behaviour. Giving it a diagonal spin.

- In this sense the information on the photon’s original polarisation is lost and so too is any informations associated with this spin.

- Quantum cryptography uses photons to transmit a key. Once the key is transmitted coding and encoding using the normal method can take place.

- The goal is to thwart attempts to eavesdrop on the message. In cryptography an **eavesdropper** is referred to as **eve.**

- Quantum cryptography is the first technique which safeguards against passive interception If she does Bob and Alice will know.

**Problems:**

- Distance: Can only transmit photons about 150km because of interference. A photon’s spin can be changed when it bounces off other particles. A lot of work is going into making photon repeaters that don’t affect spin.

- Temperature: Repeater process only currently works at very low temperatures.

- Entanglement: Photons can come to depend on another undergoing some particle reactions and their states become entangled. Entangled pairs always have the opposite spin of eachother can affect encryption.

- Weather: Fiber optic cables contract and expand with temperature, weather conditions severely affect how far we can transmit electrons.

## Hashing

- Hashing is generating a value or values from a string of text using a mathematical function.

- As opposed to encryption, hashing is a one way process that scrambles data.

- Every time we apply a hash to data we will get the same result but knowing the has does not give us the data.

- Hashing can be used to speed up searching a database.

- If you store an index of hashes, users can search for the shorter hash rather than the longer data.

- Hashing can be used as a digital fingerprint to verify the integrity of files

- When you download files from websites they will give you the computed hash for that file from different algorithms.

- After downloading, a utility can generate a hash from this file and this can be compared to the original hash to verify authenticity.

- Hashing can be used to make password storage more secure.

- When a user chooses a password, it is subjected to a hashing algorithm that transforms it into a fixed length hash value.

- The password itself never leaves the users client, instead it is the hash that the server receives and stores.

- The next time the user logs on the password is rehashed and compared to the stored value.

- The hashing algorithm is such that the hash value cannot be used to regenerate the password so if the hacker gets hold of the database it is of no use to the hacker.

- Brute forcing has got better and now certain hashes have become unsafe.

**Possible Issues**

- Data-collisions: the algorithm has to ensure there is a low risk of different files producing the same hash.

- Calculation Time: there is a computational overhead for hashing and if this is too high it may discourage usage.

- Size: the algorithm has to produce hashes smaller than the original file size otherwise there is no point using it to make comparisons.

- To make hashed passwords more secure a technique can be used called **adding salt.**

- The salt is a random string appended to new passwords before hashing.

- This makes the hash value different even for the same password.

- The salt is stored along the hash value. To check the password, the salt is used again to check the new hash against the original.

**Digital Signatures**

- Proves to the receiver that the document or data comes from the authorised sender.

- Proves the received information is unaltered.

- The sender cannot deny having signed the information (for contracts).

- Digital signatures are used to encrypt the information and to create a check-sum of the information being sent.

- This check-sum is sent with the encrypted information.

- At the receiving end, a check-sum of the data is calculated and compared with the received check-sum. If they do not match, the information has been tampered with.

- Digital signatures are commonly used to authenticate their downloaded installation files such as Windows update or anti-virus update.

**Digital Certificate**

- An digital certificate is used to verify the trustworthiness of a website ie it is the real website and not a fake.

- A digital certificate will not prove the content of the site has not been altered/hacked but at least the visitor is using the real server/

- Usually the website pays for a certificate from a certified authority. The issues certificate is used to manage the encrypted dialogue between server and browser under https protocol.

**Fast Searching**

- Hashing can be used to speed up searching a database,

- Hashes are short. If you store an index of hashes, users can search for the hash rather than long sections of data. This makes search algorithms much more efficient.