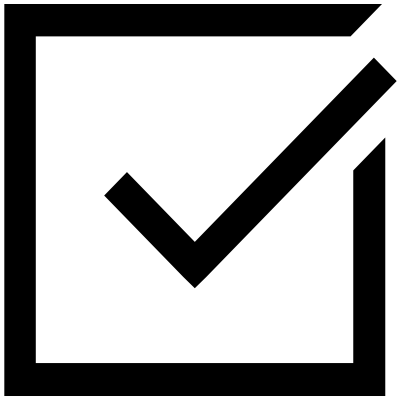
COMPUTER PRINCIPLES FOR PROGRAMMERS

File Compression and Backup

Quiz



News of the Week

Agenda

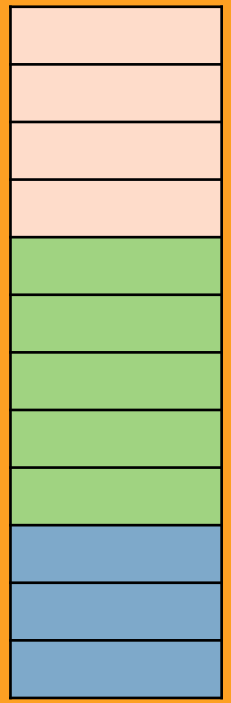
Lecture:

1. What, Why, and How of “File Compression”
   … effect on data transfer
   … drawbacks
   … overview of formats
   … Lossless vs Lossy
2. What, Why, and How of “Backup”
   … types of backups
   … backup media

Agenda (Cont'd)

- Activity:
  1. Explore File Compression
  2. Compress various native file formats to a ZIP archive and compare the results
  3. Upload files to Blackboard to demonstrate a network backup

What is “File Compression?”



4 pink
5 green
3 blue

What is “File Compression?”

- storing a file’s data in “less space” by “minimizing redundancy” in data content
- An archive is a collection of folders and files stored in one file, e.g. *filename*.ZIP
  - Files are usually compressed (but not necessarily)
- SSH (Secure SHell) can compress data: asynch.
- VoIP must do this in real-time, synchronously.

Why use File Compression?

- Sending data takes bandwidth and I/O time
  - …to a backup device
    - Tape & Optical discs – mostly obsolete
    - NAS – Network Attached Storage (local or remote intranet)
    - USB – removable drives (ad hoc, user level)
  - …to another computer
    - Via FTP or transfer to Cloud Storage over Internet
- Encrypt off-site data for security
  - compression software has a password encrypt option

The effect of File Compression on Data Transfer

Time (uncompressed file)

- 1MByte file plus ~3% TCP overhead = 8.6Mbits
- Sent in 0.345 – 2 seconds on 25Mbps network
- Sent to 30,000 users: 2.875 – 16.6 **hours**

Size (uncompressed file)

- 30,000 × ~1MB files = 30G**Bytes**
- 1G**bit** network takes 4 Minutes 30 Seconds transferring one file but will saturate network
  much more time to send 30,000 individual files

The effect of File Compression on Data Transfer (Cont'd)

- File compression rate of text to ~35% of original size translates to savings of

Time (compressed file)

- From worst case of 16.6 to 5.8 **hours**

Size (compressed file)

- 30,000 × ~1MB files = 30GB to 15~24GB
- 4:24 to 2:12~3:31 mm:ss

How File Compression works

Compression combines:

- **matching and replacement of duplicate strings with pointers.**
  - Lempel–Ziv–Welch (LZW) compression (1984)
- **replacing symbols with new, weighted symbols based on frequency of use.**
  - Huffman coding (1952)
  - David A. Huffman was a Ph.D student at MIT

How Compression Works

- Here is an old quote from Vangie Beal:

  Data compression is particularly useful in communications because it enables devices to transmit or store the same amount of data in fewer bits. There are a variety of data compression techniques, but only a few have been standardized. The CCITT has defined a standard data compression technique for transmitting and a compression standard for data communications through modems. In addition, there are file compression formats, such as ARC and ZIP.

- This quote contains 449 characters.

How Compression Works (cont'd)

- Replace " compression " with "♠". The text becomes:

  Data ♠is particularly useful in communications because it enables devices to transmit or store the same amount of data in fewer bits.There are a variety of data ♠techniques, but only a few have been standardized. The CCITT has defined a standard data ♠technique for transmitting and a ♠standard for data communications through modems. In addition, there are file ♠formats, such as ARC and ZIP.

- Including the dictionary "♠compression ", the total size is now 406 characters, 90.4% of 449.
- Compression algorithms build a token/string dictionary

How Compression Works (cont'd)

- With more pattern matching and a bigger dictionary, our quote becomes (♠compression , ♣here are , ♦communications , ♥data , ☺standard , ☼transmit , ☯technique .)

  ♥♠*is particularly useful in* ♦*because it enables devices to* ☼ *or store the same amount of* ♥*in fewer bits. T*♣*a variety of* ♥♠☯*s, but only a few have been* ☺*ized. The CCITT has defined a* ☺ ♥♠☯ *for* ☼*ting faxes and a* ♠☺ *for* ♥♦*through modems. In addition, t*♣ *file* ♠*formats, such as ARC and ZIP.*

- Including the dictionary, the total size is now 363 characters or 81% of original size. The compression advantage increases with the length of the text, i.e. more pattern matches.

Overview of some Compression File Formats

Data

ZIP, 7z, RAR
.docx, .pptx, .xlsx
StuffIt (mac OS),
.tar.gz (*nix)

Images

GIF, JPG,
PNG, TIFF

ZIP

The standard.

Music

MP3, MP4, AAC,
MQA,
OGG, FLAC

Video

MPG, MP4, DIVX,
XVID, MOV, AVI

What is “Lossless” Vs. “Lossy” Compression?

Lossless: contains all original data with redundancies removed.
Lossy: drops *things-you-won’t-notice* (you hope) from original data

Lossy vs. Lossless compression

Lossy (GIF), 45.4kb	**Lossless**, 941kb	Lossy (JPG), 25.4kb

Drawbacks to Compression

- **Time**: compressing data requires CPU cycles & network latency
  - Fast CPUs and plentiful RAM make this issue moot for PCs but not servers
  - De/compression during transmission increases network latency
- **Space**: archived files must be unarchived & uncompressed to use
- **Integrity**: any corruption can cause loss of entire archive
  - if we lose the dictionary, the file is completely unreadable.
  - Solid or multi-volume archives can be lost with even minor data corruption. *Test your archives to confirm integrity.*
- **Recoverability**: the Lossy sacrifice is reduced quality
  - original quality is lost after Lossy compression (hope you don't notice)

Three characteristics define a Backup

A **copy** in a **geographically separate location** that is **platform independent**.

Why do we need backups?

"Failure is not an option." - Apollo 13

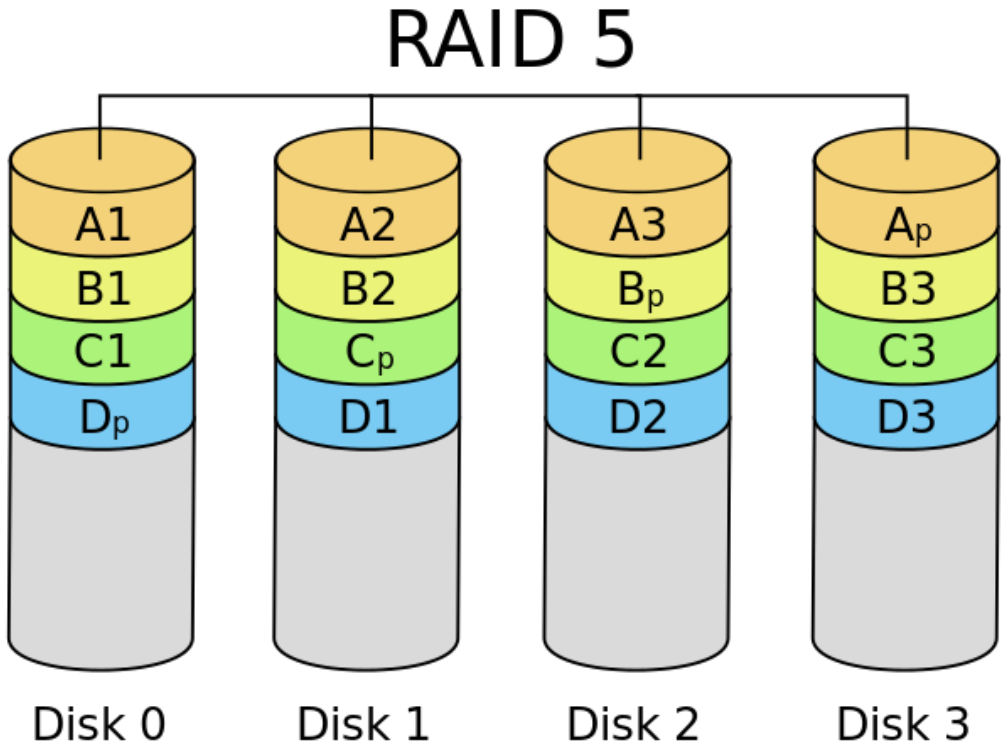
- Hardware failure: Drives fail. Power surges happen.
- Accidental deletion by users: This includes IT people.

2/3 to 3/4 of all data loss

- Hard failures
  - Mechanical: defect or wear. Catastrophy: fire, flood, theft, loss.
- Soft failures
  - software malfunction, malware / ransomware infection, SQL injection attacks, cloud provider's business failure, user errors other than accidental deletion

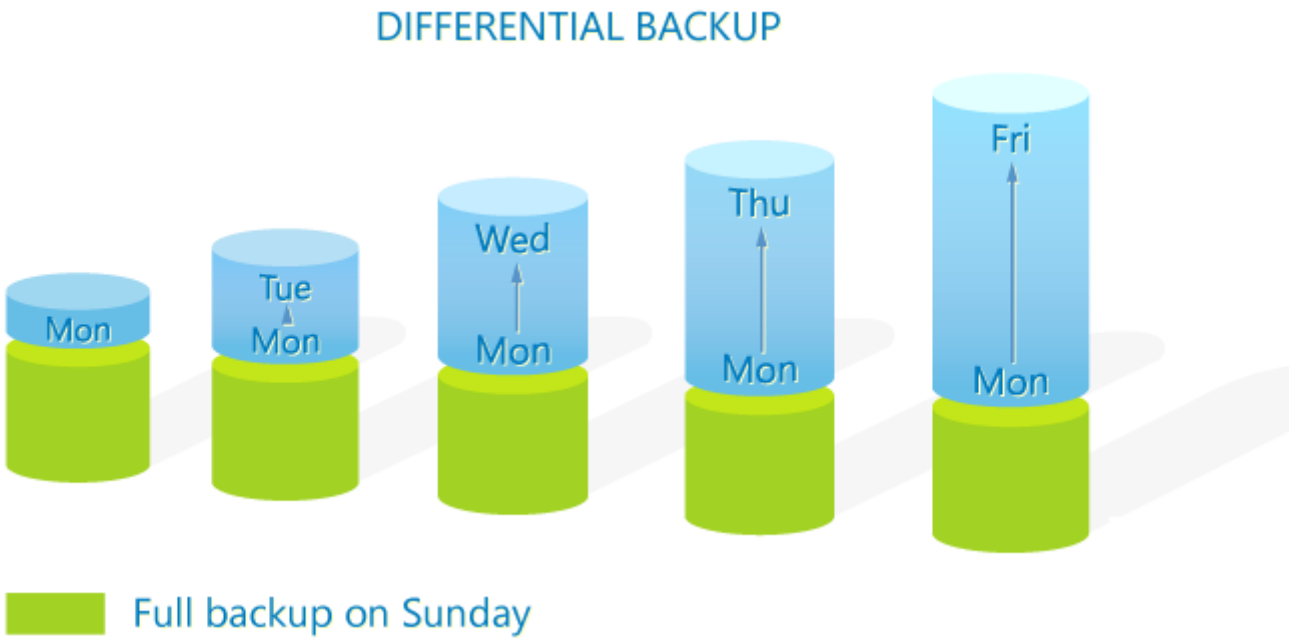
Avoiding hardware failure

- Most servers use RAID-5 to avoid disk failure
  - **Redundant Array of Independent Disks**
- Need one extra drive
- Get increased read/write performance



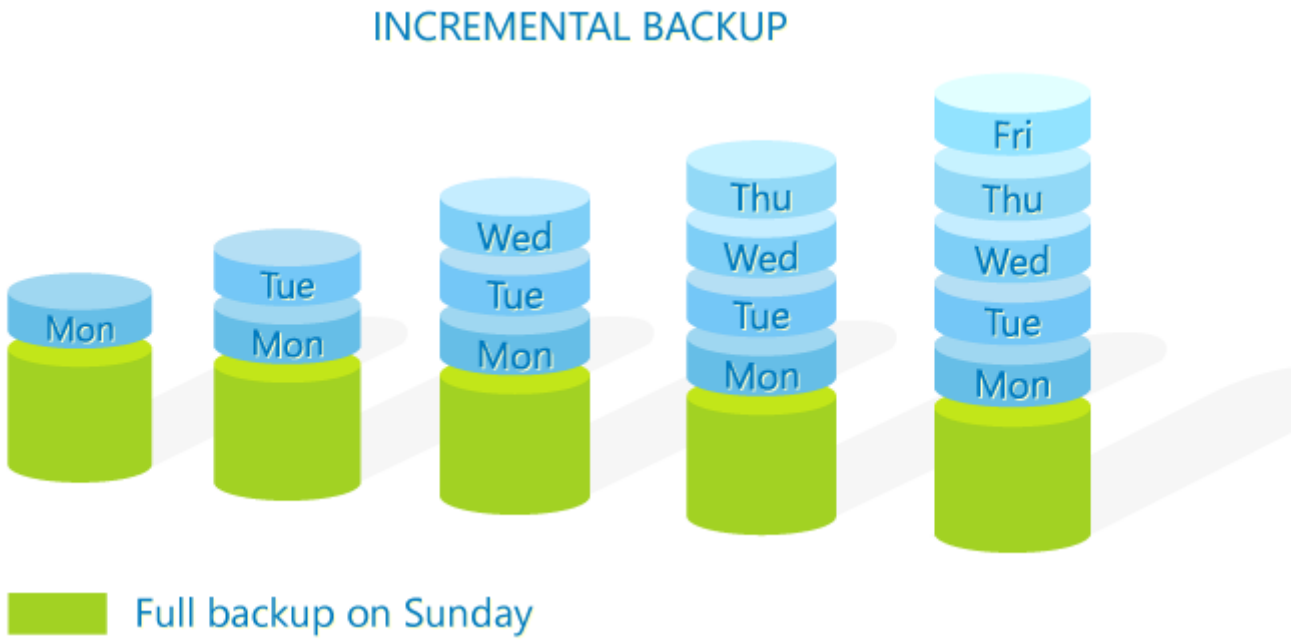
Classic File Backup Types/Strategy

Full + **Differential** (only files changed since the Full backup)



Classic File Backup Types/Strategy

Full + **Incremental** (only files changed since last backup)



Classic File Backup Schedule

- Month-end
  - Full backup, retained for 2 years
- Weekend (Sunday)
  - Differential backup, retained for 2 months
  - Differences since last Full
- Daily (Monday – Saturday)
  - Incremental backup, retained for 2 weeks
  - Differences since last backup of any type

Cloud backup options

Many network based backup options are available

- GoodSync or Bareos to devices, NAS, or cloud
- eazyBackup, Storagepipe, Sync cloud backup services
- Initial Full backup is slow, Incremental is fast. Restore?
- Options for file versioning (AKA file generations)
- Two-way synchronization is not backup for two reasons:
  - Synchronization is platform inter-dependent, not independent.
  - A file on one system does not have a "copy" on other systems, the same file **co-exists** on all synchronized systems.

Recover lost files or previous versions

Windows File History, Apple's macOS Time Machine

- Automatic copying of files to external or network drive
  - *If drive is always connected*, it is not a backup, just a copy; it is **neither geographically separate nor platform independent.**
- Historical versions of user files maintained. Easy to restore.
- Must configure and test to ensure backup of all user folders.

Windows Recycle Bin and macOS Trash can

- Only good for *oops!* and short-term recovery.
- The bin/trash is not a reliable copy much less a backup.

3-2-1 Backup Checklist

- **3 copies** *(changes to active file do not change the copies)*
  - 1 active file on your machine, 1 local copy, 1 remote backup
- **2 different formats/platforms** *(platform independence)*
  - External drive is platform independent **only when not plugged in**
  - One-way backup to cloud (not two-way sync)
- **1 off-site backup** *(geographically separate location)*
  - Cloud storage different from your cloud IaaS, PaaS, SaaS provider
  - rotating external drives from home to office

The final word on backups...

Backups do not matter.

Only RESTORE matters.

The near loss of Toy Story 2

NOTES

…not on the quiz but here for further information and explanation.

What is “Lossless” Vs. “Lossy” Compression?

- These are two basic types of compression.
- *ZIP*, *TIFF (*Tagged Image File Format), *FLAC (Free Lossless Audio Codec),* and other general file compression routines are considered *lossless* compression.
- The original data is completely encoded; compression reduces redundant data (e.g. a large blank space in a TIFF image or a long noiseless passage in FLAC audio).

What is “Lossless” Vs. “Lossy” Compression? (Cont’d)

- *JPG*, *MPG*, *MP3*, *GIF*, and other formats use *lossy* compression.
- These formats, in order to achieve compression, remove data from the source file:
  - *GIF* images *limit the number of colours* in an image to 256 per pixel but preserve detail with lossless data compression similar to ZIP files. Useful for sharp-edged line art.
  - *JPG* images *delete colour and fine detail information* to achieve compression with acceptable reproduction of photographs.
  - *MP3s simplify the sound waves of audio.*
- A file with Lossy compression *can never be returned to its original, complete state.*
- *The amount of loss or compression can be adjusted;* a developer decides *how much compression* can be applied while retaining enough useful *quality* of the content for the intended purpose. E.g. high compression for JPG thumbnail images but little compression for large, zoomable images.

Overview of some Compression File Formats

- ZIP:
  - The most popular general-purpose compression archive.
  - Supported on virtually all platforms from mainframes to PCs.
  - Includes features such as encryption using password protection.
- RAR, 7z, TAR, StuffIt:
  - Proprietary general-purpose compression file formats with incremental improvements over Zip but with the loss of standardized support.
  - use different algorithms, with various benefits and uses.
  - Some are designed for different operating systems (StuffIt for Mac, TAR for *nix--TapeARchive).

Overview of some Compression File Formats (Cont'd)

- GIF, JPG, PNG, TIFF:
  - These are compression formats used by the **graphics industry.**
  - GIF and JPG are lossy formats and cannot be uncompressed to the original source data.
- MP3, MP4, OGG, FLAC:
  - These are compression formats used by the **sound engineering and music industry.**
  - These are lossy formats (except FLAC - Fully Lossless Audio Codec) and cannot be uncompressed to original source data.

Overview of some Compression File Formats (Cont'd)

- MPG, MP4, DIVX, XVID, MOV, AVI:
  - These compression formats are used by the **video industry.**
  - These are all lossy formats.
  - These will often mix compression algorithms from audio and image. technologies.

What is a “Backup” and why do we need backups? (Cont’d)

- Backup is the “procedure for making extra copies” of data “in case the original is lost or damaged and must be restored.” The procedure includes storing the copy in a geographically separate location which is platform independent from the original file and host system.
- Having a backup will allow you to recover from lost, broken or stolen hardware, and from your own accidental deletions.
- You should be in the habit of backing up user created files on your laptop or PC. OS and apps can be restored from their original software providers or a system Restore Point but user created data can only be restored from backups.

When & How to run your Backup

- Automatically:
  - Performed by continuously running backup software that constantly monitors for file changes. Used for Full and Incremental strategy.
- Scheduled:
  - system operator or backup software runs a backup at specific times, such as overnight, when it has the least impact on business operations or at critical business times such as at accounting month/year end. Used for Full, Differential, and Incremental strategy.
- Manual Backup:
  - a user performs backups at their own convenience. Not a *strategy*.
  - It is the least effective method (what if you forget to do it?), but it's better than no backup at all!

Locations of Backup Media

- Local:
  - copies files to a drive in use by the system.
  - fastest and most convenient, but if the computer is lost or malfunctions, so goes the data! Just having a **copy** *is not* a backup.
  - Local copies may be made to reduce downtime. The copies are then moved to External media or transmitted which is a slower process.
- External:
  - Copies files to External/Portable/Flash Drive.
    i.e. a device which can be disconnected from the computer.
  - It is a backup when the platform independent device is taken off-site.

Locations of Backup Media (Cont'd)

- Network:
  - Back up files to the cloud (Google Drive, OneDrive, Dropbox, iCloud)
  - It is a slower option for large backup. Cost effective communications bandwidth has significantly less throughput than writing data to a directly attached device.
- The best location depends on the type of work you're doing, the volatility of the data (how quickly it changes), the volume of data, the backup window (available downtime), security considerations, and the speed/availability of restoration.