Contents

1 Version Control (`git`)

1.1 Overview

`git` is a version control system for software development, and is arguably the most important part of software construction. There are two main things that `git` maintains:

An object database: A repository of objects that records the history of your project

An index (cache): Records the future[1] of the project.

1.2 Getting Started

There are two main ways to start a git repository: `git init TARGET_DIRECTORY_HERE` and `git clone TARGET_REPOSITORY_HERE`

`git init TARGET_DIRECTORY_HERE` initializes an empty project inside the target directory (current directory if not specified) with a `.git` folder. This is less common, as a lot of people don't start a project from scratch.

`git clone TARGET_REPOSITORY_HERE` clones an existing repository, creating a directory on your computer containing a copty of all of the files in that repository with the `.git` folder inside that directory.

Note: When cloning a repository, it is possible to clone from a device.

Note: `git` will remember where you're cloning from; that is, if you run

```
git clone REMOTE_REPOSITORY
git clone ./REMOTE_REPOSITORY
```

`git` will identify that the second clone was from a device, whereas the first clone was from a remote location.

When working with `git`, it is important to remember that remote-to-local repositories are a downstream structure; that is, cloning from a remote repository sends a repository "downstream" to your device.

1.3 The Repository

What do you put inside your repository?

Stuff you change by hand

What should you **NOT** put inside your repository?

Automatically generated files (e.g. `node_modules`)

Stuff that isn't portable/shouldn't be portable (e.g. `.env.local`)

`.gitignore`: By default, `git` creates this file, which will tell `git` to automatically ignore file/type(s) that are specified inside `.gitignore`. This file is a very important one, especially to keep your repository clean and portable.

ASIDE: Shorthand for Commit ID's

`COMMIT_ID^`: The commit before `COMMIT_ID`

`COMMIT_ID~n`: The `HEAD` - n^{th} commit

`COMMIT_ID^!`: Same as `COMMIT_ID^..COMMIT_ID`

`COMMIT_ID..COMMIT_ID`: Range of commits (`start`, `end`]

[1] Future: plans for the future of the project, immediate or long-term

1.4 Managing the Repository

The following subsections will cover common git commands that are used to manage the repository.

Note: All of these commands are called under the assumption that you're in the current repository folder.

1.4.1 State

This set of commands gives information of the state of the repository.

`git status`

`git status`: Tells you the current status of your repository. Mainly, it will list all files that have been added, modified, or deleted relative to your last commit.

`git ls-files`

`git ls-files`: Lists all working files managed by `git` to `stdout`. Files that are not tracked by git will not show up on this list(hence why we do not just use `ls`).

`git blame`

`git blame`: Returns a line-by-line history of a specified file in a specified commit (`HEAD` if not specified) with the author and timestamp of each line.

`git diff`

`git diff COMMIT_A..COMMIT_B`: Takes a `diff` of two commits and prints to `stdout` **(See `git log` for navigating the Terminal output)**.

`git grep PATTERN`

`git grep PATTERN`: Same as doing `grep PATTERN $(git ls-files)` **(See `grep`)**

`git log`

`git log [OPTIONS] (start-point..end-point]`: Prints the commit history between `start-point` exclusive to `end-point` inclusive in reverse-time order (new → old). Prints the entire commit history from the first commit to the most recent commit if no `start-point, end-point` are specified.

Options

`-n`: Look at the HEAD - n^{th} commit

`--decorate[...]`: Format git log output with specified parameters **(See HW4)**

Navigating `git log` in the Terminal

`/PATTERN`: Searches for a pattern in the output.

`n` and `N`: Goes to the next and previous n^{th} occurrence respectively

`q`: Exits the log output

`SHIFT-g`: Scrolls to the very bottom of the log output

ASIDE: Using `git log`

`git log` is commonly piped into other commands such as `wc` **(See `Shell Commands`)**

`git log` outputs both the committer and author of a commit. While often times they are the same person, it may be that they are not. This is more apparent in big open source projects with controlled/reviewed commits. The person with repository access will be listed as the committer, while the person who wrote the code will be listed as the author.

1.4.2 Pushing Upstream

This set of commands relates to pushing upstream to the central repository, mainly staging, committing, and pushing.

`git clean`

`git clean`: Removes all untracked files from the repository.

`git add`

`git add FILE`: Stages a file to commit. If the file was previously untracked, `git` will now track the file.

`git rm`

`git rm FILE`: Removes file as well as untracks the file that was removed. This is equivalent to doing `rm FILE` **(See Shell Commands)** followed by `git add FILE`.

Options

`-f, --force`: Forcefully remove file and ignore any warnings

`-r`: Recursively delete a directory and all of its contents

`git reset`

`git reset [OPTIONS]`: Unstages all modified files.

Options

`--soft`: Only reset `HEAD`

`--hard`: Reset `HEAD`, the index, and working tree

`git commit`

`git commit [OPTIONS]`: Creates a commit with all of your staged files and allows for a commit message.

Commit Semantics A commit message should explain **why** they are adding to the repository, not what they are contributing. A commit message should have the following format:

```
brief summary here

* more
* details
* here
```

Options

`-m`: Write a commit message inline

`--amend`: Amend a previous commit.

Note: Amending should be done sparingly and never in big open source projects to avoid confusion.

`git push`

`git push [OPTIONS]`: Pushes all commit(s) from your local repository upstream into the central repository.

Options

`-u, --set-upstream`: Set upstream branch to push to

`--atomic`: Request atomic transaction on remote side

1.4.3 Pulling Downstream

This set of commands relate to pulling from the upstream repository, mainly fetching and pulling.

`git fetch`

`git fetch [OPTIONS] [BRANCH]`: Fetches metadata from a remote branch, `origin` if not specified. Note that all of the working files in the local repository remain unchanged. This is generally a safer way to update your local repository with the lastest metadata since it does not change any working files.

Options

`--all`: Fetch information from all remotes

`--atomic`: Request atomic transaction on remote side

`git pull`

`git pull [OPTIONS] [BRANCH]`: Pulls metadata from a remote branch, `origin` if not specified. Note that all of the working files in the local repository will be updated to match the upstream repository. If the local branch is behind the remote, the local branch will fast forward by default. If there are divergent branches, use either the `--rebase` or `--no-rebase` option to resolve conflicts. `git pull` will fail if there is no specified method of resolving conflicts since `git` is conservative[2]. `git pull` is equivalent to `git fetch` followed by `git merge` or `git rebase` depending on default configurations.

Options

`--all`: Fetch information from all remotes

`--atomic`: Request atomic transaction on remote side

1.4.4 Branch Manipulation

1.4.5 Overview

A `branch` is a lightweight **moveable** pointer[2] to a commit. By default, when creating a repository, there is only one branch, `main/master`. By default, when creating a new branch, `git` will branch off of the current branch. `git` is a tree structure, meaning it must be a `DAG` in order to work.

`git branch`

`git branch [OPTIONS] [BRANCH]`: Lists all of the repository's local branches.

Options

`-d`: Delete the branch `BRANCH`

`-D`: Delete the branch `BRANCH` without warning

`-m`: Renames a branch from `A` to `B`

`git checkout`

`git checkout [OPTIONS] BRANCH`: Changes all of the working files to be identical to the ones in the specified `BRANCH`. Alternatively, `git switch BRANCH` is similar but has a few minor differences.
When checking out, `git` is conservative[3] and will prevent a checkout if you have uncommitted or untracked working files.

Options

`-f, --force`: Force a checkout and ignore any warnings

`-b`: Create a new branch `BRANCH` and start it at the `start-point` of the `main` branch

`-B`: Resets `BRANCH` to a specified `start-point` if the branch exists, same as `-b` otherwise

[2]Pointer: In `git`, `HEAD` is a reference variable that points to the tip of the current working branch.

[3]Conservative: `git` will warn you if you have uncommitted or untracked files when performing any actions that mutate your working files.

git merge

`git merge [BRANCH]`: Merges branch `BRANCH` into the current branch. This creates a graphical commit history.

```
         A---B---C topic
        /         \
   D---E---F---G---H master
```

git rebase

`git rebase [BRANCH]`: Reapplies commits atop a branch tip. This creates a linear history rather than a `DAG`.
Given a commit history,

```
        A---B---C topic
       /
   D---E---A'---F master
```

Running `git rebase master` will produce

```
                B'---C' topic
               /
   D---E---A'---F master
```

git bisect

`git bisect` runs a binary search to find the first bad version.

```
git bisect start
git bisect bad (current version)
git bisect good VERSION
```

Note: `git bisect` might not work on merged commit histories.

1.5 Extraneous `git` Features

1.5.1 Tags

Tags essentially label commits, and are created by running the command `git tag COMMIT_ID`. There are various types of tags: plain, annotated, and signed tags. They are located in `refs/tags`. It is worth noting that branches and tags can be the same names.

Plain

Plain tags are the literally just giving names to commits. There is no metadata stored.

Annotated

Annotated tags store metadata and can be created by running the command `git tag -a TAGNAME -m "MESSAGE" COMMIT_ID`.

Signed

Signed tags are for security, and have cryptographic authentication. They can be created by running the command `git tag -s`

1.5.2 Submodules

Submodules in git are used to "point" to another project. It contains the commit ID's within the other project and is used for version stability. To update submodules, run the command `git submodule foreach git pull origin master`.

1.5.3 Stashing

Stashes are implemented with a stack, and are used for switching branches. `git stash push/pop` will push/pop your modified working files onto/off the stack respectively. `git stash list` will list all of your stacks. If you want to be avant garde like Eggert, you can instead do:

```
git diff > mychanges.diff
patch -p1 < mychanges.diff
```

1.6 Communicating Between Developers

There are multiple ways to communicate between developers:

GUI enthusiasts: Share a repository and use pull requests via something like Github

CLI enjoyers: Email patches back and forth:

`git format-patch A..B`

`git send-email`

`git am FILE` (automatic merge)

2 Build Tools

Who is the audience for these build tools?

Developers: Write the source code for the software

Builders: Compile source code for a particular platform

Distributers: Ship programs to users in the form of distributions

Installers/Configurers: Users that install and use the programs

2.1 `make`

Once developers are done writing their code, they want to help builders compile it. In order to do so, they must provide metainformation about the source code and all of its dependencies. One easy way to do this is by writing a metaprogram that will automatically implement these build instructions. In simple programs, rather than a metaprogram, a README is used. Otherwise, we can write a simple script (commonly labeled `build.sh` or `setup.py`). Here's what a sample script may look lke:

```
gcc -c a.c
gcc -c b.c
gcc -c c.c
gcc -c a.o b.o c.o -o foo
```

2.1.1 Flaws/Fixes

There are a couple downsides to this approach

(a) Maintaining this file can be too time consuming/get confusing

(b) Rebuilding after small changes is expensive

(c) Not scalable

(d) It's slow (missing parallelism)

How do we fix these issues? We can't fix all of these issues, but we can use a separate build tool rather than write our own script via `Makefiles`.

2.1.2 Makefiles

Makefiles are similar to shellscripts but are more efficient: they only rebuild what is necessary. A sample Makefile may look like:

```
a.o: a.c
    gcc -c a.c
b.o: b.c
    gcc -c b.c
c.o: c.c
    gcc -c c.c
foo: a.o b.o c.o
    gcc a.o b.o c.o -o foo
```

`make` will determine what needs to be rebuilt by looking at file timestamps. Additionally, we can run jobs in parallel with `make -j 10`This solves (a), but this approach also creates new problems/doesn't fix old problems.

(i) Clock skew: Different machines might differ in their exact system time and if they're operating on the same set of files, it's possible that one system writes a timestamp that is ahead of another system's time or the program file generated by make is older than the edited timestamp.

(ii) Missing/Extra Dependencies may cause the program to break//rebuild unnecessarily.

2.2 Syntax

2.2.1 $

`$`: Expands a variable

```
OBJ = a.o b.o c.o
foo: $(OBJ)
    gcc $(OBJ) -o foo
```

is equivalent to

```
foo: a.o b.o c.o
    gcc a.o b.o c.o -o foo
```

2.2.2 $@

`$@`: Expands to the rule name.

```
foo: a.o b.o c.o
    gcc a.o b.o c.o -o $@
```

is equivalent to

```
foo: a.o b.o c.o
    gcc a.o b.o c.o -o foo
```

2.2.3 Rules and Recipes

Rules have the following syntax:

```
TARGET: DEPENDENCIES
    RECIPE
```

Note: Recipes are shellscripts. Furthermore, `make` is a thin layer around the shell.

3 C (The Superior Language)

C is the predecessor to C++, so it is missing a lot of 'features' that C++ has. Some of these are:

(a) STL

(b) Classes and Objects

  (i) Polymorphism (`foo(int& a)` and `foo(bool a)`)
  (ii) Inheritance (`class Dog: public Animal`)
  (iii) Encapsulation (`private`)

(c) Namespace Control

(d) Explicit use of `static` to create singular instances

(e) Exception Handling

(f) Memory Management: `new` and `delete` (wrappers for `malloc()` and `free()` respectively)

(g) `cin`, `cout`, `<< >>`

(h) Function Overloading

3.1 Architecture of a C Environment

Compilation is broken up into different stages:

(1) Preprocessing (`gcc -E foo.c` $\implies$ `foo.i`)

(2) Conversion to ASM (`gcc -S foo.i` $\implies$ `foo.s`)

(3) Create Object Files (`gcc -c foo.s` $\implies$ `foo.o`)

(4) Linking (`gcc *.o` $\implies$ `a.out`)

Note: At (3), the object files have holes in them. We need to resolve this by linking all of the `.o` files which produces a single executable which will cut and paste all of the `.o` files in the correct place.
Note: The preprocessing phase is usually omitted by higher level languages (e.g. Python). Essentially, preprocessing allows for conditional compilation via `#ifdef, #ifndef, #endif`, and other macros.

4 Debugging

Debugging a program serves two main purposes:

(1) Correctness: Verifying that the expected output matches the actual output

(2) Performance: Change code to optimize for hardware/better performance

In real-time systems (car brakes), correctness and performance are indistinguishable, since they are dependent on each other. In general, try to avoid using a debugger (Eggert's words not mine). Below are some alternatives to debugging that should be tried before busting out a debugger.

(1) Print Statements (`cout`, `printf()`, ...): To track variable states

(2) `time`: To measure the efficiency of the program (and deduce any timing issues)

(3) `ps -ef`: Prints all active processes

(4) `ps -efjt`: Similar to `-ef`, but in tree form

(5) `top`: List of top-consuming processes (by CPU %)

(6) `kill`: Kills a process

(7) `strace ./a.out foo`: Logs to `stderr` all system calls

Note: Most often, (1) and (2) are most commonly labeled under developer tools, while (3) and its subitems are labeled under operation team tools.

ASIDE: System Calls

System calls are special commands executed by the OS kernel, which lives right atop the hardware level. This is more of an OS topic but it still proves relevant in this course, especially since we talk about the `gcc` compiler. System calls are special since only the OS kernal can actually execute these instructions. Most other applications must **ask** the OS kernel to execute the syscall.

4.1 `valgrind`

`valgrind` is a debugging tool mainly used to detect memory-related bugs and to log **all** instructions a program executes.
`valigrind ./a.out foo`
`valgrind` isn't perfect, but it does help against many trivial memory-related bugs such as bad references. `valgrind` will catch

```
char *p = NULL;
*p = 'x';
```

but won't catch

```
char a [10000];
char *p = &a[10000];
*p = 'x';
```

since `valgrind` won't do trivial boundary checks by default. **See HW 5** for more information.

ASIDE: The Stack

`gcc -fstack-protector` is there for a reason: to prevent malicious people from injecting code into the program's instruction list, overflowing the buffer, and taking control.

4.2 `gcc`

`gcc [OPTIONS] [FILE]` has many options to help you debug. Here is an important one:
`-fstack-protector`: Protects against stack overflow errors by inserting a canary right around stack boundaries. If the canary is not a predictable value, the stack was corrupted, so the program will crash gracefully. Note that this won't always work since there are ways to get around this and still cause stack overflow errors.

4.2.1 Profiling

`gcc --coverage` will profile your program, creating a temperature graph by injecting code into your program like

```
if(x < 0)
   counter[19246]++;
   f();
```

and will output `counter` to an output file (`counter` is the profile). Note that profiling is input-dependent.

Profiling is done to find bugs with cold functions (a.k.a why are the cold?). However, this is also test-case dependent, since if functions are labeled cold, it might be because your test cases never touch them.

4.2.2 Static Checking

Static checking prevents your code from compiling if it fails a static check/assert and are used to document your code and assumptions. They have the format
`static_assert(E)`, where `E` is a constant expression. So, asserts like

```
int f(int n) {
    static_assert(0 < n);
}
```

will not work, since `n` is not a constant variable.

4.2.3 Warning Flags

`-Wall`: `gcc` will turn on all "useful" warning flags

`-Wcomment`: Catches bad comments like `/* bad /* comment */`

`-Wparentheses`: Catches potential arithmetic errors like `return a << b + c` (`+` has higher operator precedence than `<<`)

`-Waddress`: Warns about using addresses that are probably wrong. e.g. Consider the following:

```
char* p = f(x);
if (p == "abc")
    return 27;
```

`-Wstrict-aliasing`: Warns against "bad" casts. e.g. Consider the following:

```
long l = -27;
int *p = (int*)&l;
*p = 0;
```

`-Wmaybe-uninitialized`: Warns if you're using potentially uninitialized variables.

4.2.4 Optimization

`gcc` has an optimization flag that will trade compile time for faster executables.
`gcc -O#` (0-4, 2 being the most common) will determine the level of optimization.

4.2.5 Overview

The two most common ways `gcc` optimizes your source code is by caching in registers and executing out of order. This makes your code harder to debug when you run it, since what you see is not always what you wrote in the source code.

4.2.6 `-O#` Alternative: `-flto`

`gcc -flto`: An alternative to the plain `-O#` flag, we have `gcc -flto`, or File Time Link Optimization. This will put a copy of the source code into all of the `.o` files and will optimize the entire program at once, with all of the modules linked. This way, there is more opportunity for optimization. The main downside to this approach is that compile times are even slower.

4.2.7 Built-In Compiler Functions

Below are a list of common functions to help optimize or debug your source code:

(a) `__builtin.unreachable()`: Tells the compiler that if the program ever reaches `unreachable()`, then behavior is undefined. This allows for further optimizations. e.g. Consider the following:

```
if(x < 0)
    __builtin.unreachable();
return x / 16;
```

Since the compiler knows that `x` *should* never be negative, it can use the bitshift operation `x >> 4` to optimize.

(b) `__attribute__(ATTR)`: Advice to the compiler (can be ignored). Does not change the program. This allows for further, nuanced optimization. e.g. Consider the following:

```
#ifdef __GNUC__
#define __attribute__(x)
#endif
```

The above code will disable the attribute if compiled with a non-gcc compiler.

4.2.8 Attributes

```
charbuf[1000]__attribute__((aligned(8)));
```

`aligned(x)` makes sure that `charbuf` has an address with a multiple of `x`, where `x` is a power of 2. This is to maximize the number of CPU cache hits. Since RAM is divided into cache boundaries, the machine will cache (usually) 64-bytes of memory on the CPU. Doing `aligned(x)` will (try to) ensure that the array fits into the cache's 64-byte boundaries.

```
void func(void) __attribute__((cold))
```

`(cold)/(hot)` labels a function either cold or hot, respectively. A cold function is one that is rarely executed, whereas a hot function is one that is executed frequently. The motive behind this is so that the instruction pointer does not have to jump around everywhere and can execute (relatively) sequentially.

```
instruction pointer
V
------------------------
| hot | program | cold |
------------------------
```

This is how the compiler will order your code using attributes.

```
int hash(char*, ptrdiff_t) __attribute__((pure, access(read_only, 1)));
int a = hash(p, 27);
int b = hash(p, 27);
```

`pure` means that there is no user-visible storage. In this case, `a` must equal `b`.

```
int square(int)__attribute__((const));
```

`const` means that the value is both `pure` and does not depend on user-visible storage. In C, `pure` ≡ `[[reproduceable]]` and `const` ≡ `[[unsequenced]]`

```
void *myalloc(ptrdiff_t) __attribute__((alloc_size(1), malloc(free, 1), returns_nonnull))
```

4.2.9 Runtime Checking

`-fsanitize=undefined`: Runtime check for overflows

`-fsanitize=address`: Crash if bad pointers are used

`-fsanitize=leak`: Check for memory leaks

`-fsanitize=thread`: Check for race conditions

ASIDE: `unsigned`

`unsigned` is a disaster for one very specific reason. Let `x` be an unsigned integer. Now consider:

```
if (x <= -1)
```

This statement will always evaluate to true because `x` is unsigned. Logically however, this makes no sense.

4.3 Debugging: Using gdb

There are a couple prerequisites before using gdb:

(1) Stabalize the failure (make sure it consistently breaks)

(2) Locate the source of failure (point of failure)

(3) Optionally, `gcc -g` will put information such as names of local variables to make debugging easier.

```
1. (gdb) set cwd /usr
2. (gdb) set env TZ American/Chicago
3. (gdb) set disable randomization on(default)/off
4. (gdb) r -c foo < bar >baz
5. (gdb) r
```

(gdb) `attach PID`: Takes over process id

(gdb) `b foo`: Breakpoint at `foo`

(gdb) `info break`: Lists breakpoints

(gdb) `del #`: Delete breakpoint `#`

(gdb) `step, s`: Step to the next line

(gdb) `stepi`: Step into the next machine instruction

(gdb) `next, n`: Step over function calls

(gdb) `cont, c`: Continue execution

(gdb) `fin`: Finish current function

(gdb) `bt`: Backtrace (examine current state)

(gdb) `p E`: Print the value of the expression `E`

(gdb) `target TARGET`: Target a specified architecture

(gdb) `reverse continue, rc`: Reverse execution

(gdb) `checkpoint`: Will output a unique id of the program state

(gdb) `restart ID`: Restarts execution starting from `ID`

(gdb) `watch E`: Pause execution when `E` changes

4.3.1 gdb with Optimization

When debugging it is important to remember that the executable may behave differently than what is written in the source code due to **optimization (See Optimization)**.

Out-of-Order Execution

Consider the following source code:

```
(1) q = a / b;
(2) r = a % b;
```

may turn into

```
r = a % b;
q = a / b;
```

since, in a lot of architectures, the instruction `idivq` will calculate both the division and modulo. This is due to the "as-if" rule: The compiler can generate any code whose behavior is "as if" it did the obvious. Therefore, one method of debugging is to do the following:

```
gcc foo.c
gdb a.out
[debug]
gcc -O2 foo.c
[run]
```

The problem with this is that the optimizer may be buggy (unlikely), or the optimizer exposed a bug that wasn't caught when debugging (more likely).

4.3.2 Finding Bugs

Suppose

```
start
...
bug triggered*
...
...
failure
```

How do you find the point of failure(`*`)?
In small programs or programs with easy test cases, we can:

Come up with a reproducable test case

Make sure the program doesn't take too long to execute

Rerun the program until you find it

For larger programs, we can use gdb's **Reverse Execution** to find the bug(s).

Reverse Execution

gdb will start executing the program backwards. Note that this is a very expensive process since gdb has to cache all program states. To efficiently use `gdb rc`, we can use commands like `checkpoint`, `restart` and `watch`. Catchpoints stop the program if it throws an error (similar to a try/catch block). ***Note:*** `gdb watch` is so cool that a lot of architectures have hardware support for `watch`, meaning it's fast. On x86-64, you can `watch` up to 4 memory locations.

4.3.3 Review

Try not to do it; that is, write good test cases

Test cases > source code: Test-Driven Development - Write test cases before coding the corresponding part

Use a better platform: e.g. Subscript errors? C++ $\rightarrow$ Rust or Java

Defensive Programming

Assume other devs are useless

Runtime checking

Trace/log what the program does along the way (helps debugging later)

Assertions

Exception handlers (try/catch)

Barricades: Middleware to take in any data and only pass through "safe" data into the program

5 `git` Internals

5.1 Preface: Atomicity and SHA-1

An atomic operation only has two states: not executed or executed e.g. `cd`. Non-atomic operations such as `cp` are logical since it is possible to be in the middle of writing a file when execution stops (unexpectedly). `git` uses many atomic operations to keep the working tree clean and to prevent corrupting the repository. For example, `git commit` is built atop atomic operations because it would not be good to only have half of a commit.

The SHA-1 checksum is the hash function that `git` uses to create commit ids and object hashes. Though it has been cracked, `git` still uses it because:

> The probability of collisions is $\frac{1}{2^n}$, where n is 160 in this case
>
> Finding a byestring to match a given hash is expensive ($O(2^n)$) (SHA-1 is a one-way hash)
>
> Finding collisions is expensive ($O(2^n)$)

5.2 Overview

`git` is like an application-specific "file system" (because it was built by file system designers). It is built atop an ordinary file system and has many similar issues that file systems have. `git` is split up into to parts: plumbing and porcelain. The plumbing part deals with the internals such as data structures and low level commands, while the porcelain part is what the user interfaces with (e.g. `git commit`).One of the main issues with `git` is distinguishing data with metadata.

5.3 .git/

Below are some of the subdirectories/files under `.git/` and their usage.

`.git/ branches/`: Legacy folder (for backwards compatability) that used to store branches

`.git/ config`: `git`'s configuration file. Analogous to a barricade **(See Debugging)**

`.git/ description`: Descriptor for the repository

`.git/ HEAD`: The pointer[4] that points to the tip of the current working branch

`.git/ hooks/`: `git`'s callbacks

`.git/ index`: Binary data structure keeping track of your commit

`.git/ info/exclude`: Contains blobs that git will ignore (like `.gitignore`, but not for working files)

`.git/ logs/`: Record your branch tips and logs changes to branches (2^{nd} order history: history of the **repository**)

`.git/ objects/`: Contains the object database with records of all objects managed by the repository

`.git/ refs/`: The references directory to store commits and tags

`.git/ packed-refs`: Condensed version of `refs`

Below are some more notes on the following contents of `.git/`

config

There is no mixing of data and metadata. That is, do not include anything in the .git folder in the repository. This leads to a very natural question: How do I share my `.git/config` file? The solution is to write a script to set up the `config` file and put instructions in a README.

[4] See *Pointer* in **Managing the Repository: Branch Manipulation**

objects/

Objects in `git` are identified via a 40-digit hexadecimal (160-bit) checksum[5]. The objects folder will store all objects managed by `git`. The subdirectories (e.g. `objects/0f`) contains the first two digits, while the file descriptor inside contains the remaining 38 digits. This was done to meet the storage requirements at the time. Nowadays, it's still formatted this way for backwards compatability.

refs/

`refs/heads/BRANCH_NAME`
Points to the last **local** commit ID of `BRANCH_NAME`. e.g. `.../main` will contain the most recent **local** commit ID of `main`.

`./remotes/origin/HEAD`
Contains the relative file path of where HEAD points. e.g. `ref: refs/remotes/origin/main`.

`refs/remotes/origin/BRANCH_NAME`
Points to the last **remote** commit ID of `BRANCH_NAME`. e.g. `.../main` will contain the most recent **remote** commit ID of `main`.

`refs/tags/TAG`
Contains all of the repository's tags

5.4 Representing Objects in `git`

Objects in `git` are not files. Rather, they are a blob containing a hashed byte-string. The following subsections will manually build a `commit` object.

5.4.1 Working Files → `blob`

The command `git hash-object FILENAME -w` will create an object with the 40-digit SHA-1 checksum (hash) for its file descriptor. Note that if two files have the exact same contents, then `git hash-object` will return the same 40-digit checksum.
Note: `git cat-file -p/t HASH` will print either the contents or type of the object with the hash `HASH` respectively.

5.4.2 `blob` → `tree`

The command `git update-index --add --cacheinfo <MODE> <HASH> <FILENAME>` will add the object to the index, where `MODE` is the type of object (e.g. `blob` = 100644). The first 3 digits is the filetype (100 = regular file) while the next 3 is the octal representation of permissions (644 = o+rw, ag+r).
The command `git write-tree` will create a tree object using the current index.

5.4.3 The commit Object

A `commit` object contains the following:

tree

commit message

author + timestamp

committer + timestamp

parent commit(s)

Note: `BRANCH_NAME` is a commit object. More generally, branches deal with commit objects. Additionally, `git` compresses objects.

[5]The checksum is calculated via the SHA-1 hash, and is used to avoid collisions.

5.5 Compression

5.5.1 Overview

Compression is the process of reducing file size while preserving as much data as possible. Many techniques are used to compress data, and the various compression algorithms are application-specific. There are trade offs to compressing: CPU time to compress/decompress, % compressed/decompressed, and RAM usage are all inversely related.

Problems
If any data gets corrupted during compression or decompression, neither algorithms works and any remaining data is now suspect.

5.5.2 Huffman Coding

The algorithm for huffman coding is very straightforward:

(1) Sort character frequency in non-decreasing order

(2) Take the least two likely symbols with the smallest weights and combine them, adding their weights

(3) Delete the two individual symbols from the list and add the new combined symbol(s) to the list

(4) Repeat (2) and (3) until there is only one node left

Adaptive Huffman Coding is a variation of the huffman tree, in which the decompressor builds the Huffman tree as it receives data, updating the tree in real-time.

5.5.3 Dictionary Compression

The Dictionary Compression algorithm is similar to a sliding window algorithm, and is as follows:

(1) Create a dictionary of byte string

(2) Send one byte string at a time, sending the offset and size between a recurrence (if there is one) and the first occurrence (if within the sliding window) instead.

(3) Repeat until End of File

5.5.4 `git` Compression

To compress objects, `git` uses `zlib/gzip` which use both **Huffman Coding** and **Dictionary Compression** (e.g. Raw Data → Dictionary Compression → Huffman Coding).

6 A 1h 20m Aside: Character Encodings

6.1 Overview

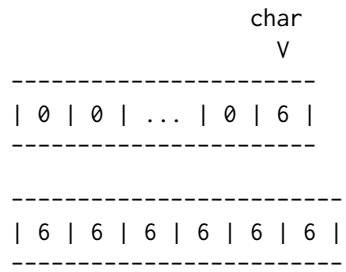
In computers, there is no such thing as a character. Computers only store numbers, so characters are just mapped integers. An easy example is the C/++ character. In C/++, the character 'x' can be represented as 'x', 120, or '\170'. Therefore, characters are just an individual symbol that corresponds to a small integer.

Corollary

A character string is a sequence of characters. From above, we have that a character is just an integer. So, it follows that a character string is a sequence of integers.

6.2 Dark Ages

In 1960, There were only 64-bit character encodings: A-Z, 0-9, +, -, *, /, etc. There is a problem with this approach however. If, by example, the wordsize is only 24 bits, 26 bits are being wasted. A simple fix is to afix wordsizes to be 36 bits. Then, take a corresponding 36-bit word and divide it into 6 blocks, where each block is any character that can be represented with 6-bits. Below are diagrams for the 24-bit and 36-bit wordsizes respectively.



6.3 EBCDIC

In 1964, IBM System 360 (Mainframe) introduced byte addressing which separates addresses of bytes. They used 8-bit bytes and 32-bit/4-byte words. Current x86-64 machines have 512-bit registers and 64-bit words. EBCDIC expanded the character set to 8-bits.

6.3.1 Flaws/Fixes

For some reason, they did not make lowercase letters contiguous and left gaps/holes in the character encoding table. These idiots did not listen to Eggert and clearly did not follow test-driven development, since they would've made it better otherwise. This is why no one uses it anymore.

There are no fixes for this bum-ass character set. Notably, Eggert wasn't able to write a C program that did character arithmetic, so they got an F in CS35L and did not pass.

6.4 ASCII

ASCII is a 7-bit character set and is superior to EBCDIC since they listened to Eggert's request of wanting to write a C program that did character arithmetic. They use 8-bit word sizes, but the first bit is a parity bit[6]. There are 32 control characters that won't print to the console (0-31st characters on the table). Some interesting things to note is that NULL is all bits 0 (7'b0) and DEL is all bits 1 (7'b1) for historical reasons.

6.4.1 Flaws/Fixes

ASCII does not natively support other languages since it's character set is so small. The devs clacked their three braincells together and came up with ISO/IEC 8859 (why 8859 I have no idea), which was a guide for how to extend[7] ASCII to other languages.

8859- 1: Latin-1 (Western-European languages)

8859- 2: Latin-2 (Central + East-European languages)

8859- 3: Latin-3 (Southern-European languages)

8859- 4: Latin-4 (Northern-European languages)

8859- 5: Latin-5 (Cyrillic languages)

8859- 15: Latin-9 ("Fixed" Latin-1 which added some bullshit French character and minor languages, and added the euro symbol)

[6] A parity bit XORS all of the other bits for error detection

[7] These extensions were not allowed to collide with the original ASCII encodings

While these were great bandaids, these bums clearly fell asleep in Eggert's lecture on test-driven development, since these extensions are not cross-compatible. Furthermore, metadata for character encodings is required to determine which character set to use when parsing (e.g. For HTTPS, we have Content type ... charset = "ISO 8859-1" in the header). Lastly, the developers did not take into consideration Asian languages, which I can't really knock them for since Asian languages have character sets longer than my notes for this class.

6.5 Encoding for Asian Languages

Developers said "fuck it we ball" and increased to 16-bit character sets to encapsulate Asian languages like basic Chinese. In C, we cannot use `char` anymore, so we have to use `short`'s.

6.5.1 Flaws/Fixes

The problem now is that it's completely incompatible with any other character encoding schema. e.g. Something like "Hello" will be parsed (in ASCII) as

```
-------------------------------------
| 0 , 'H' | 0 , 'e' | ... | 0 , 'o' |
-------------------------------------
```

where `0` is the null-byte. Furthermore, this encoding is very obviously bloated.

To fix the incompatibility, they used multibyte characters, which had the following format:

1-byte characters for ASCII had parity bit 0

2-byte characters for others (e.g. Kanji) had parity bit 1

This encoding was called ShiftJIS and was adopted by Microsoft and ASCII[8]. These developers were big fans of the Hydra[9] because their "fix" also introduced two issues. Firstly, the file **must** be processed sequentially due to character **context**. Moreover, this schema introduced more invalid encodings.

6.6 Unicode Consortium

Unicode was an attempt to "unify" Asian languages and have a single universal charcter set for all characters and languages. There are currently 149,186 assignments. In the 1990's, the developers did not futureproof for emojis and though that a 16-bit character set would be enough.

6.6.1 Flaws

Unicode has a lot of repeat characters that are virtually identical but there were national debates over some goddamn lines and that's why we have a lot of repeat characters (most common in Asian languages). One of Eggert's favorite examples is the Latin vs. Cyrillic 'o'. They look the same but apparently there's a slight difference. I'm not going to do a `diff` of the character pixel maps so I'll take his word for it.

6.7 UTF-8

UTF-8 is upwards compatible with ASCII. Its schema is as follows:

Every multibyte sequence has only non-ASCII bytes (parity bit 1). This way, it is easy to see character boundaries

There are 3 byte types:

ASCII byte: parity bit 0

Continuation byte: parity bit 1 and 2nd bit 0. It an **never** be a leading byte.

Length + Leading bits byte: First k bytes are the length of the character

[8] ASCII was a Japanese company completely unrelated to US-ASCII (similar to how Javascript is not related to Java in any way)

[9] In Greek mythology, the Hydra was a serpentine water beast which, when one of its heads were cut off, two more would grow back in its place

UTF-8 Boundaries

```
------------
| 0XXXXXXX |
------------
```

U+0000 - U+007F

```
------------  ------------
| 110XXXXX |  | 10YYYYYY |
------------  ------------
```

U+0080 - U+07FF

```
------------  ------------  ------------
| 1110XXXX |  | 10YYYYYY |  | 10ZZZZZZ |
------------  ------------  ------------
```

U+0800 - U+FFFF

```
------------  ------------  ------------  ------------
| 11110XXX |  | 10YYYYYY |  | 10ZZZZZZ |  | 10WWWWWW |
------------  ------------  ------------  ------------
```

U+FFFF - U+10FFF

6.7.1 Flaws

No character encoding is perfect, UTF-8 included. There are gaps in UTF-8 encoding since there are multiple ways to spell characters:

`11000001 10111111`

is technically the DEL key, but these encodings were accounted for (as invalid UTF-8 encodings) since the developers did not fall asleep in Eggert's lecture on character encodings. Moreover, byte-for-byte comparisons won't work because something like `strcmp("UCLA", "UCLA");`, where the first and second UCLA's are 1-byte and 2-byte respectively, will return false. Additionally, something like

```
char *p = XXXXXX;
p[strlen(p)/2] = 0;
```

won't work in UTF-8.

More invalid UTF-8

```
| -------------
| | 10XXXXXXX |
| -------------
```

Continuation byte **must** follow length bytes

```
------------
| 111110XX |
------------
```

Max length is 4

```
------------ |
| 1110XXXX | |
------------ |
```

Length bytes must be at the start*

Note: This may be a part of a datastream that hasn't sent all of its packages over yet, so you have to be careful when checking for valid UTF-8 encoding. This is why `Barricades` are important.

One common coding convention is to use ASCII only to prevent any encoding errors.

7 Backups

According to Eggert, we backed up a total of 100 ZB[10] in the past year, roughly 90% of which is dulicate data and roughly 50% in the cloud. Backups very clearly dominate storage, and there is a cost for backups (global warming, apparently). Do we need all of these backups? If you look at M152A computers, you'll know that to a certain group, 93 backups (with extremely similar names) are necessary for a singular lab.

7.1 Overview

Backups are a snapshot of file contents (with metadata for each file). There are two types of backups: abstract and concrete.

Abstract: Each file is a byte string (byte sequence with separate byte strings for data, metadata, etc.). This means it's dependent on OS but it isn't wasteful since you only copy over exactly what you need.
Concrete: Abstract the actual data into blobs[11] and instead, copy the blocks in the underlying device. This means it's independent of the OS and captures the exact state of the device, but it could potentially be wasteful since in practice, the device might contain bloat.

Regardless of methodology, backups address a multitude of problems:

(1) Data loss

(2) Hardware failure

(3) Tracking history

(4) Accidentally trashing a working copy because you didn't follow Eggert's Best Practices™

(5) Corrupted drives (Hardware failure but with some chest hair)

(6) Security (ransomware)

Backups used to just be an operaton staff (Ops) problem, but they couldn't handle it so now it's a DevOps problem.

7.2 Cheaper Alternatives

(1) Simply generate less data, use compression or back up less often (who would've thought)

(2) Multiplex your backup: multiple drives backed up onto one bigger drive

(3) Incremental backups: Back up only what changes. Note that this is more fragile (but is very similar to (1))

(4) Selective backups: Determine what is worth backing up.

(5) Snapshots: **See Snapshots)**

(6) Backup to cheapder devices. e.g.

```
Flash => Disk => Optical
Main    Backup  Secondary
                 Backup
```

(7) Redundancy in devices: **(See RAID (Redundant Array of Inexpensive Disks))**

[10] 1 ZB = 10^{21}

[11] Blobs stand for "Binary Large OBjectS" and "isn't made up", which I don't really believe but whatever.

7.2.1 Incremental Backups

At the file level, each backup has a timestamp, so take a similar approach to `make` **(See `make`)** and only backup files with t' > t. Consequently, we run into the same problems as `make` like clock-skew. So, in yet another layer of abstraction, we rely on the clocks being monotonically nondecreasing. One other problem with this is that deleted files are not addressed in this schema.

Within a given file F, consider ΔF. You can do `diff -u F ΔF > t`. You now have an "edit script" that will patch a file F to ΔF by running `patch <t F`. This is good for text files.

7.2.2 Automated Data Grooming

Deduplication is the process of automatically removing data we don't need. The algorithm works as follows:

```
find all file where g == f
    for each g
        rm g
        ln f g (there is a race condition BUT ln -f f g is atomic)
```

This assumes the files are read-only, since if you now change `g`, `f` is also changed **(See Hard Links)**. To remedy this problem, we have Copy-on-Write (CoW), which will make a copy of a file if it's link count > 1, writing to the copy. The idea is to share read-only files, and make a copy for writes.

This leads to another issue: If metadata(f) ≠ metadata(g), g will lose metadata. To solve this issue, we change the definition of equality. Finally, there's the issue of not having enough storage to copy on write.

Block-level Deduplication

Let a particular file system have 8 KiB blocks. We can represent it as:

```
-----------------
|   | A  |   | A |
-----------------
```

Using block-level deduplication, there is only one copy of A. More generally, the file system will only save distinct blocks (this is default on many Linux distros). This way, we get an implicit Copy-on-Write for free. There are three main issues with this type of deduplication:

(1) Allocation: Not enough storage to copy on write

(2) Slower access time: "What's another level of indirection?" is what the devs said, laughing

(3) Reliability: If a block goes bad, you're screwed

7.3 Backups and Encryption

Reasons for encrypting backups:

(1) You don't trust your cloud provider

(2) You don't trust your operations staff (lol)

(3) Data must be encrypted for other reasons (security)

7.4 Bridge to Version Control Systems

7.4.1 Preface: Versioning and File Systems

Do applications need to know about backups?

Yes: Software like Files-11 (OpenVMS) will create viewable backup files, so when you do `ls -l`, you get something like

```
foo.c; 1
foo.c; 2
...
```

so that applications now have an API for versioning.

7.4.2 Snapshots

No: Utilize snapshots, which captures the current state of your file system in user-specified invervals. This method is used on SEASnet via a NetApp file server that runs WAFL (block-level deduplicaiton).

ASIDE

Directory size is irrelevant (it has a nice personality). Why? You can't directly read from a directory; that is, you cannot do something like `cat DIRECTORY`.

7.4.3 History

SCCS in 1972 was the first major proprietary VCS, and it worked as follows: for each source file F, $\exists$ s.F which contained the entire history of F in increasing time order as well as metadata (committer, message, etc.). This let a user read any version via a single sequential pass. However, the downside was that the cost of retrieval, at worst, was now O(size of history).

A free alternative was RCS, which was similar to SCCS, but structuerd as follows: for each working file F, $\exists$ RCS/F.v, where F.v was the history of the file in the format:

```
------------------
     Metadata
------------------
Most recent change
------------------
Reverse time order (e.g. 12 => 11)
------------------
...
------------------
2 => 1
------------------
```

One major issue with RCS was that it was a per-file VCS. The creator of RCS wasn't as smart as Linus Torvalds.

CVS (not the pharmacy) introduced commits that can address multiple files, and had a client-server model for repositories. A descendent of CVS was SVN, which was CVS on steroids.

The Linux kernal initially used: CVS $\rightarrow$ SVN $\rightarrow$ BitKeeper (proprietary software). Linus Torvalds said "fuck that I want free" so naturally, he built `git`, which hilariously ran BitKeeper out of business (they open sourced in 2016 but hardly anyone uses BitKeeper anymore).

8 A 10 min Overview of Compiler Internals

Compiled languages (like C/++) compile in multiple stages **(See C)**. The hardest part however is converting into general ASM. Compilers answer the question of "How do I turn
`a += *b[5]` into
`movq b, %rbx`
`movl 0 XX`"

Let L denote the many source languages (C/++, Python, etc.) and M denote the many architectures (x86-64, ARM, RISC-V, etc.). Do we have to write $L \cdot M$ compilers? No! Instead, we have a set of common compiler internals that take in a language $l \in L$ and translate it to a specific architecture $m \in M$, which then converts into general ASM. This way, we only need to do $c + L + M$ work, where c is a constant.

9 Software and Law

9.1 Software

Software is:

A set of instructions to a computer

A way to collaborate with other users and developers to solve problems

9.2 Law

Law is "the art of predicting judges"[12]. It can be broken up into multiple categories:

How to collaborate

How to deal with failures in collaboration

Civil/Contract/Commercial

Criminal

Constitutional

International

Admiralty (oceanic)

9.2.1 Commercial Law and Software

Back in the Dark Ages, copyrights and patents were very different. Copyrights were reserved for creative works like books, while patents were reserved for functional inventions like a urinal headrest[13]. Nowadays, the line between copyrights and patents are starting to blur due to software.

Software is used with hardware, but software would technically be copyrighted while hardware would be patented.

Trade Secrets

Trade Secrets have no expiration date, and expire when the secret is disclosed. Note that if you illegally disclose a secret, it is still legally a secret. There are agreements to keep secrets called "Trade Secret Agreements". This is more commonly referred to as an NDA, or a Non-disclosure Agreement and many companies make you sign one.

[12]This quote was authored by Paul Eggert, UCLA Senior Lecturer, in La Kretz Hall 110 on February 16, 2023
[13]Hilariously enough, this was a real, granted patent.

Trademarks

Like Trade Secrets, Trademarks don't expire until a company stops using it. The goal is to avoid customer confusion. So, if trademarks don't collide, it's ok (e.g. Apple computers and Apple Records).

Personal Data

Whenever you visit a site, websites have access to your IP address and browser fingerprint.

Copyright

Copyrights cover creative works, and protects the form, not the idea. (e.g. I can write a book about whale-catching and I wouldn't be infringing on Moby Dick's copyright). Inversely, the Public Domain is any creative work that is free to use and isn't copyrighted.

Patents

Patents cover practical works like inventions and utility. To be granted a patent, you have to apply for one, and it gets reviewed. The invention must be: novel, useful, and it has to work.

9.2.2 Infringement

Legal protection for copyright/patent holders (under civil law). Infringement penalties include damages (actual[14] or statutory[15]) and takedown notices (DMCA).

9.2.3 Technical Protection

For software, you can use SaaS (Software as a Service) or program obfuscation.

9.3 Licensing

A license is **not** a contract, but rather a grant permitting you to do something, and is often part of a contract and has strings attached. When do they come up?

Buy vs Build: Using already developed software or writing your own

Derivative works: Building off of other people's work

The different types of licenses (free $\rightarrow$ proprietary) is as follows:

Public Domain: Free use

Academic: Must give credit (e.g. MIT License)

Reciprocal: Share and share alike (e.g. GNU Public License)

Corporate: e.g. Apple, Oracle

Proprietary: Paid service

9.3.1 Dual Licenses

Products can be distributed under multiple licenses (e.g. MariaDB has a proprietary version and a free (1 yr delayed) version). The reasoning for licensing and free software is so that you are in the company's ecosystem.

9.4 Software and Laws of War

"casus belli" and "jus ad bellum" translate to "case for war" and "justification for war" respectively. Is a software attack enough justification to go to war?

[14] Actual: Calculated losses

[15] A minimum they pull out of their ass