Principles of Software Construction

Version Control in the Wild

Claire Le Goues **Bogdan Vasilescu**

Carnegie Mellon University
School of Computer Science

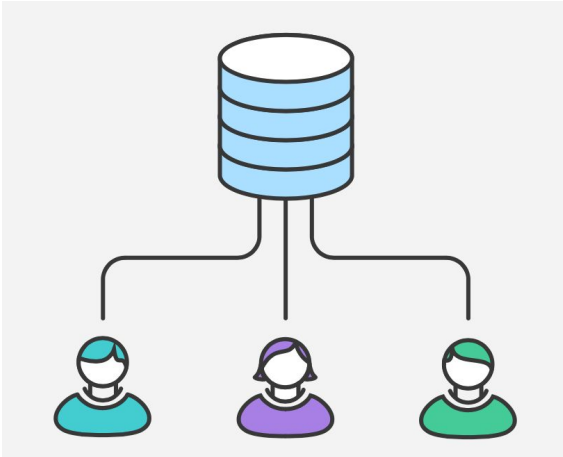
institute for SOFTWARE RESEARCH

BRANCH WORKFLOWS

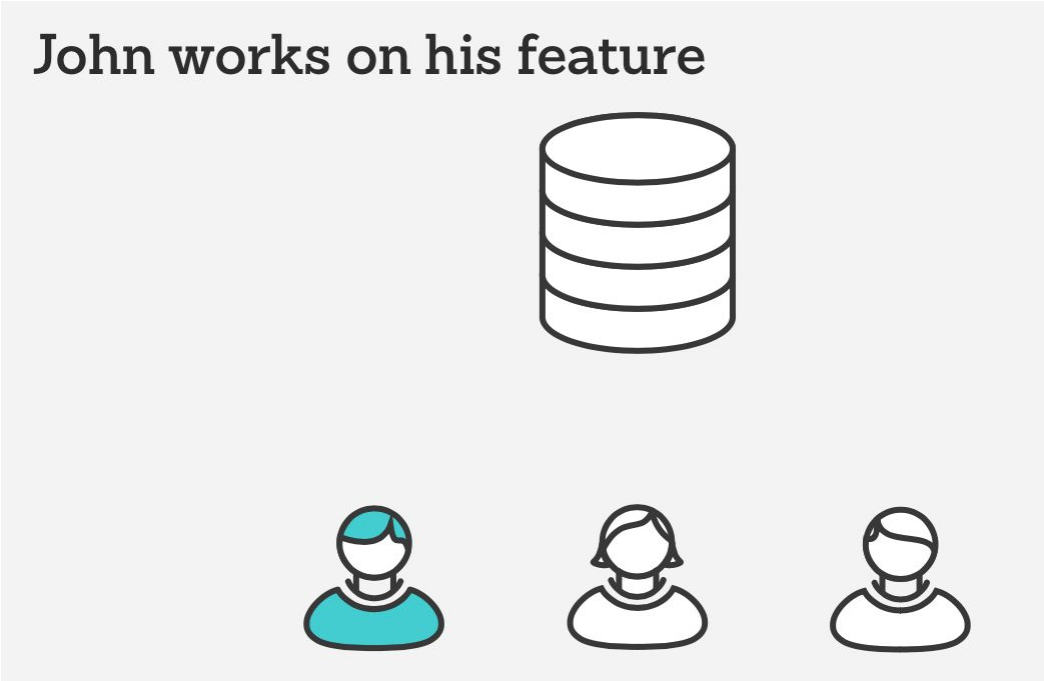
https://www.atlassian.com/git/tutorials/comparing-workflows

1. Centralized workflow

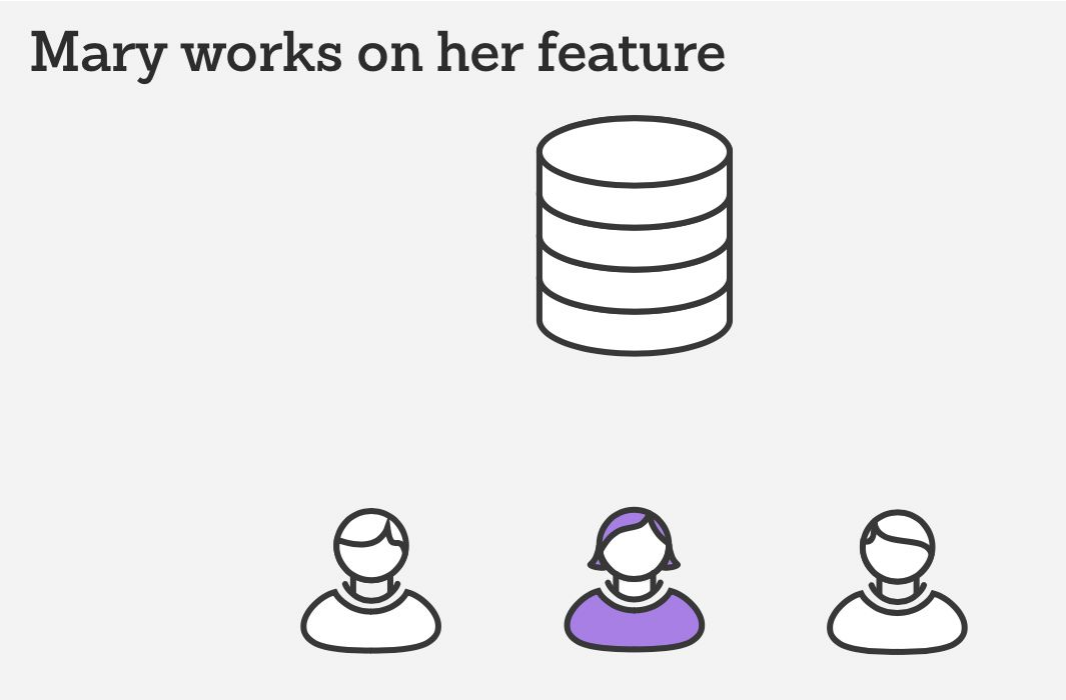
- Central repository to serve as the single point-of-entry for all changes to the project
- Default development branch is called main
  - all changes are committed into main
  - doesn’t require any other branches



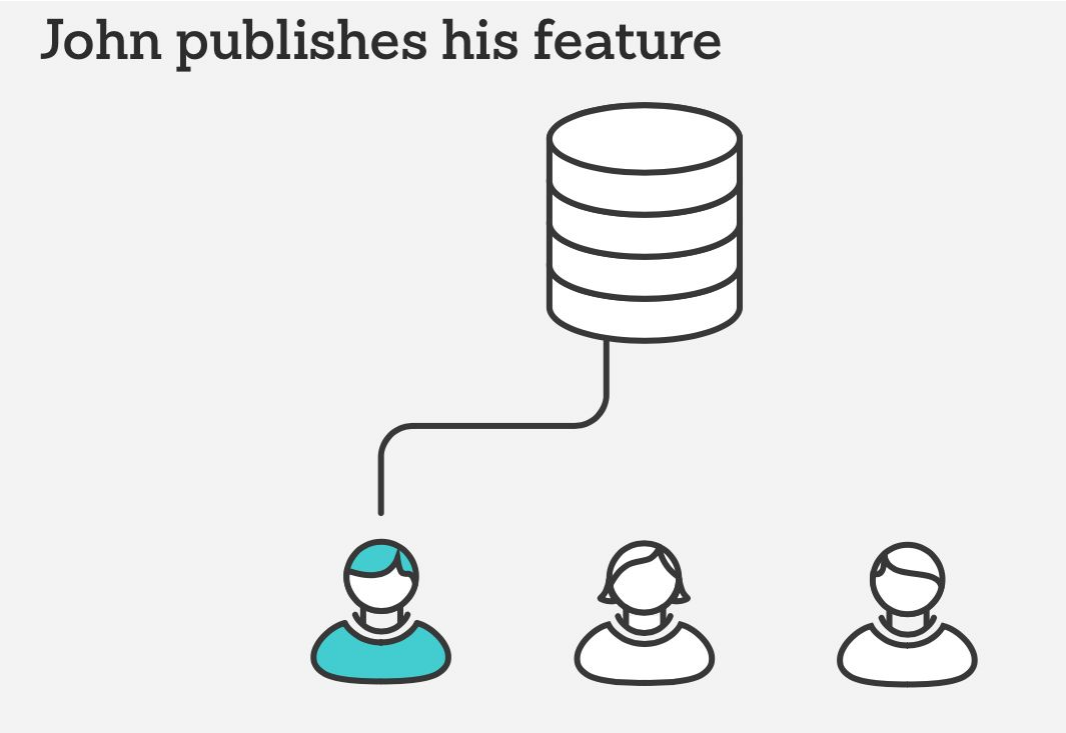
Example



Example

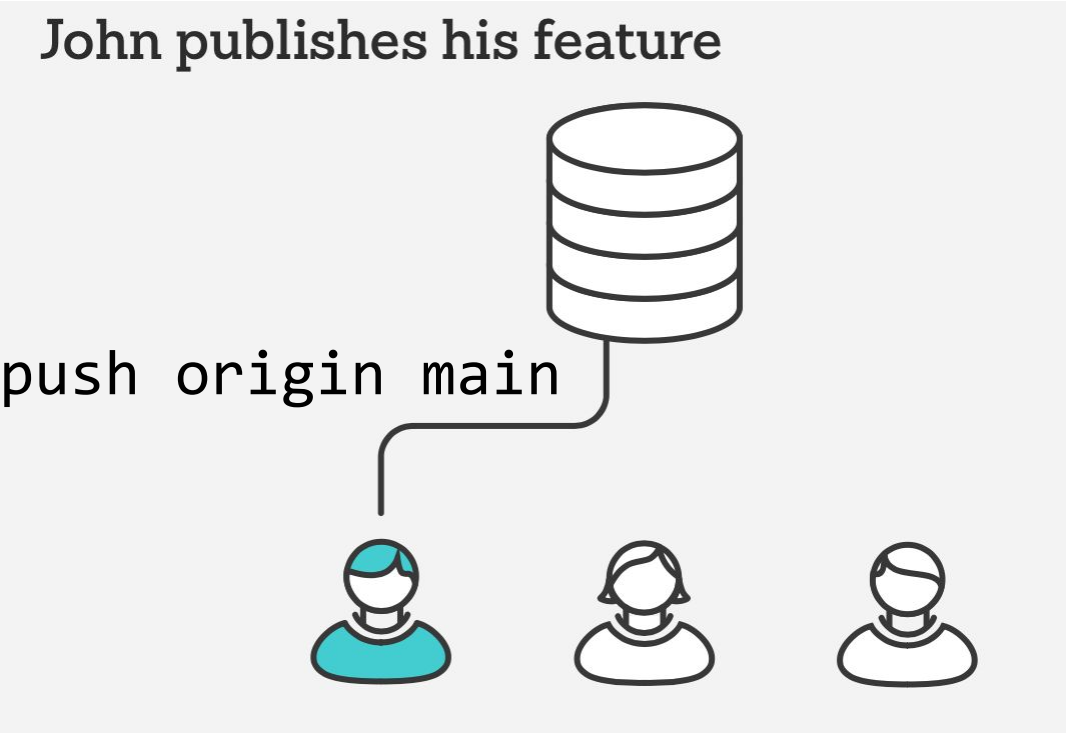


Example

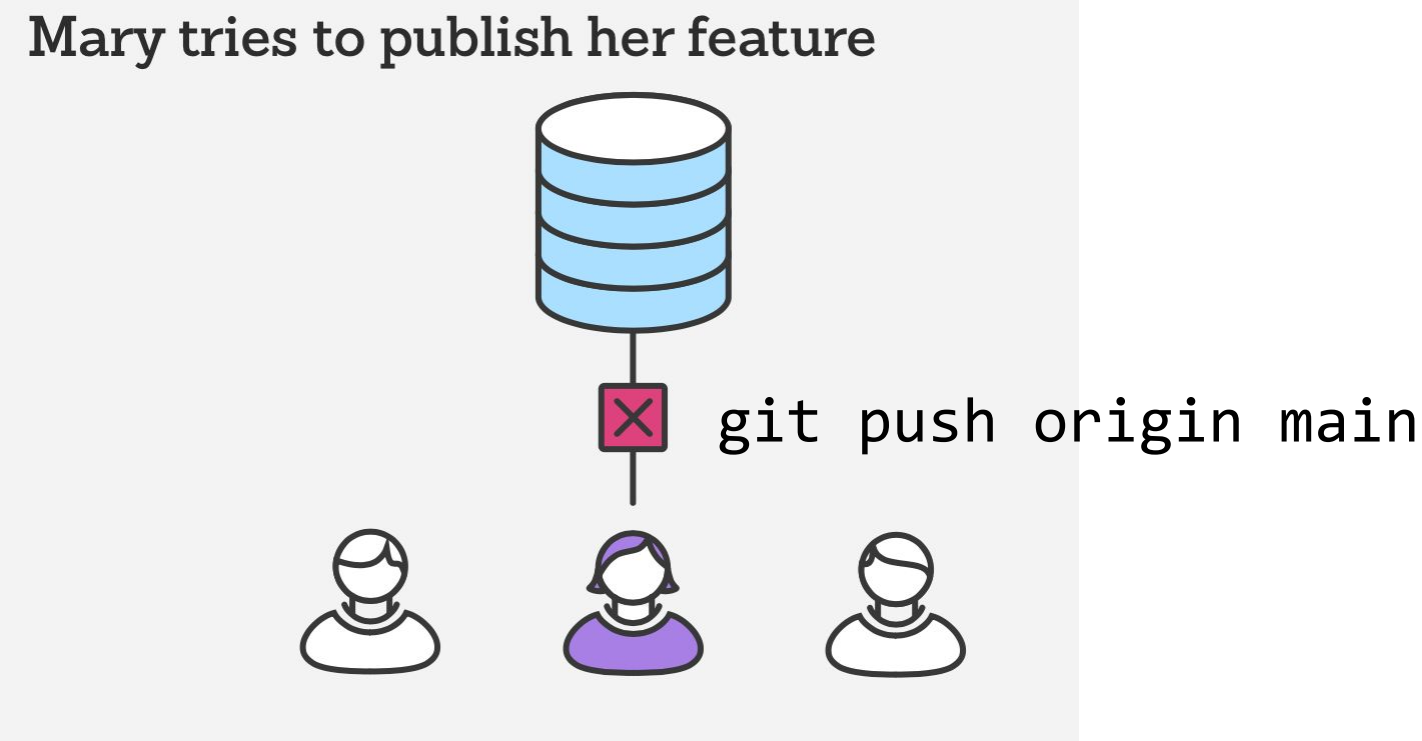


Example

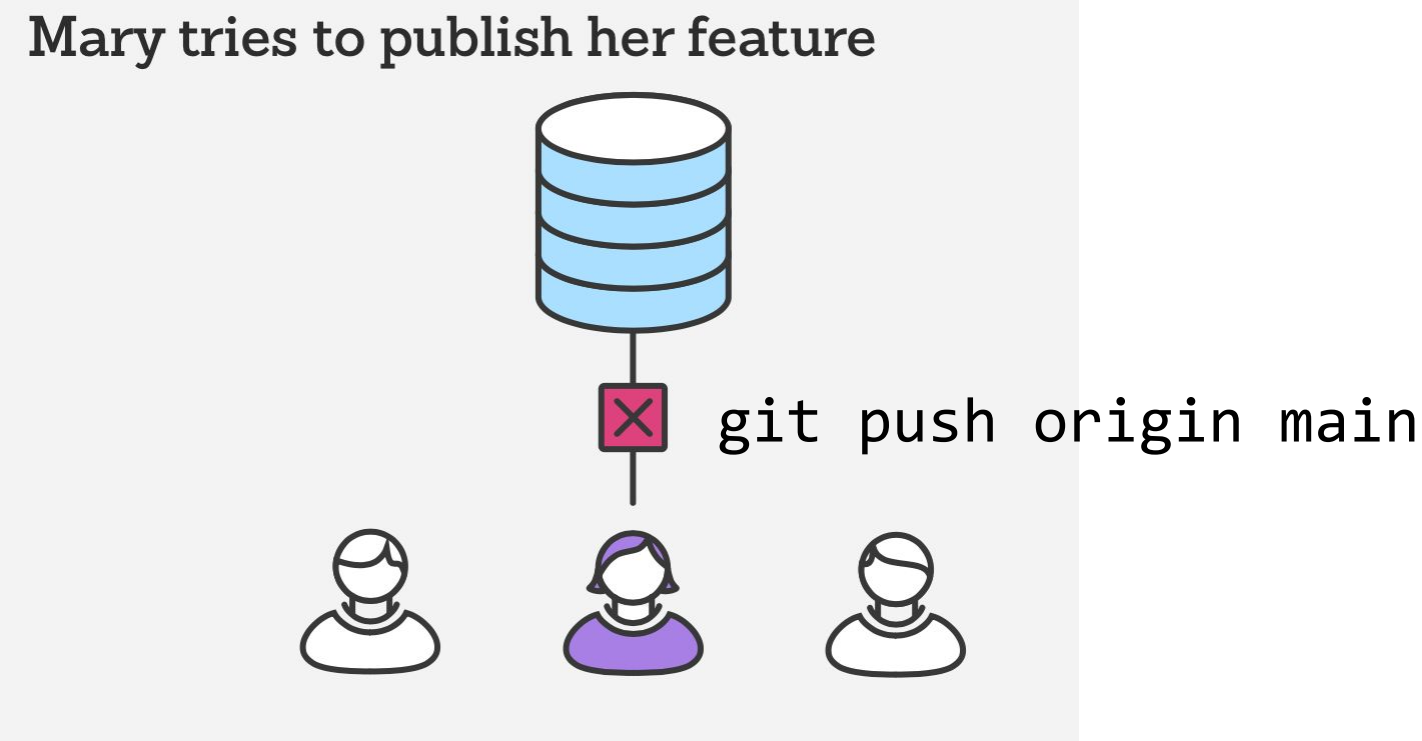
```
git push origin main
```



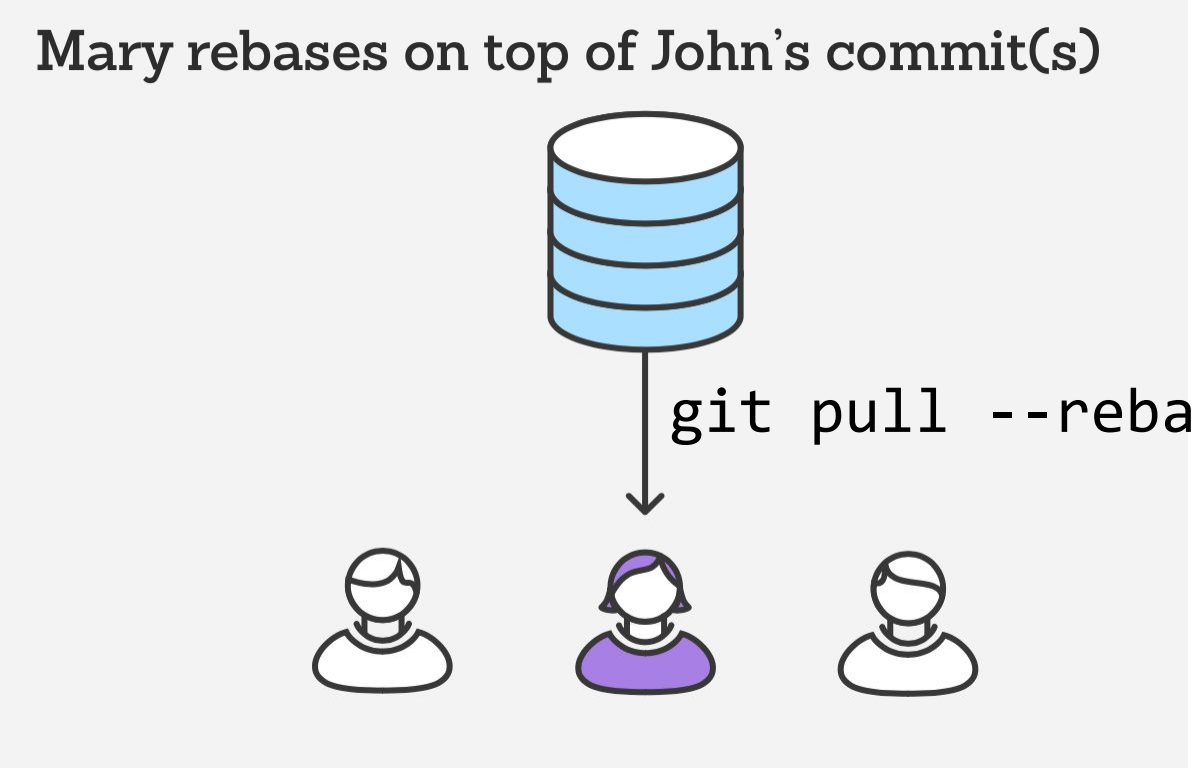
Example

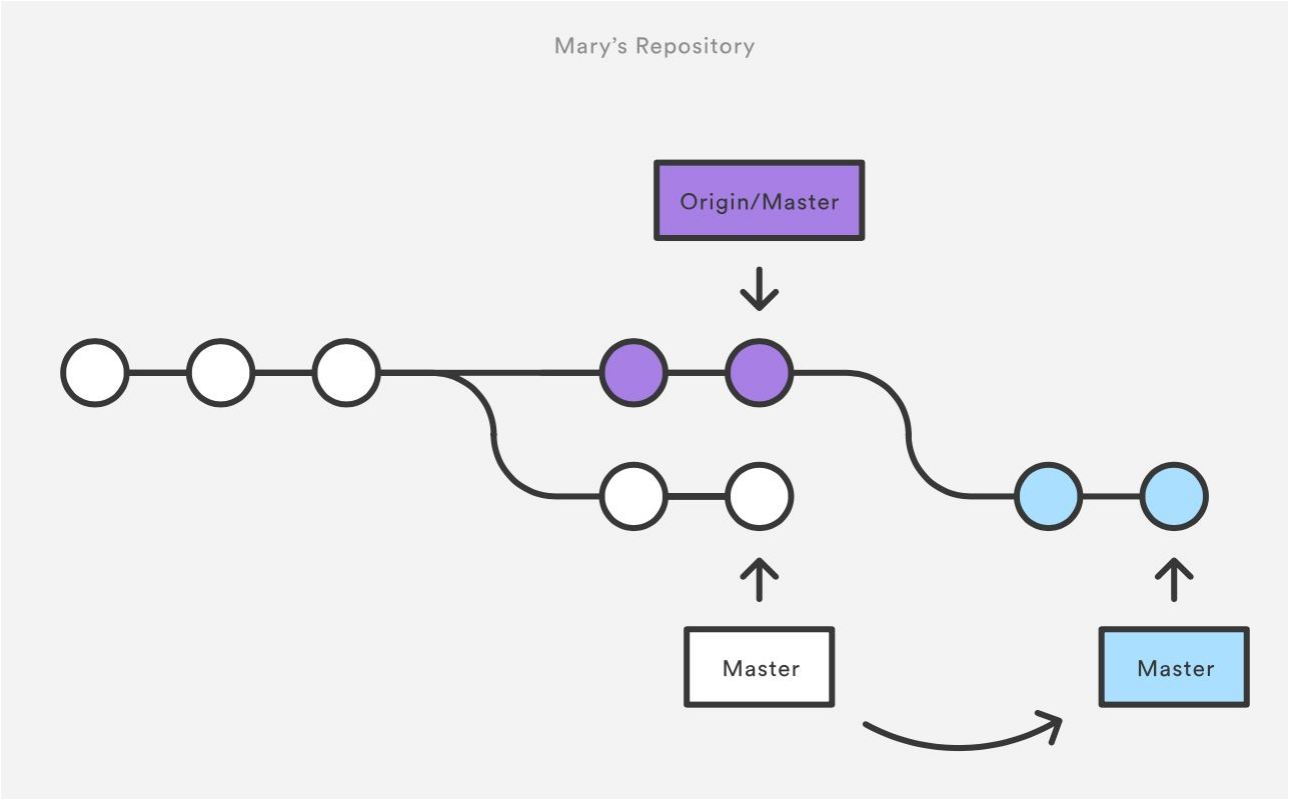


```
error: failed to push some refs to '/path/to/repo.git'
hint: Updates were rejected because the tip of your current branch is behind its
remote counterpart. Merge the remote changes (e.g. 'git pull') before pushing again.
See the 'Note about fast-forwards' in 'git push --help' for details.
```

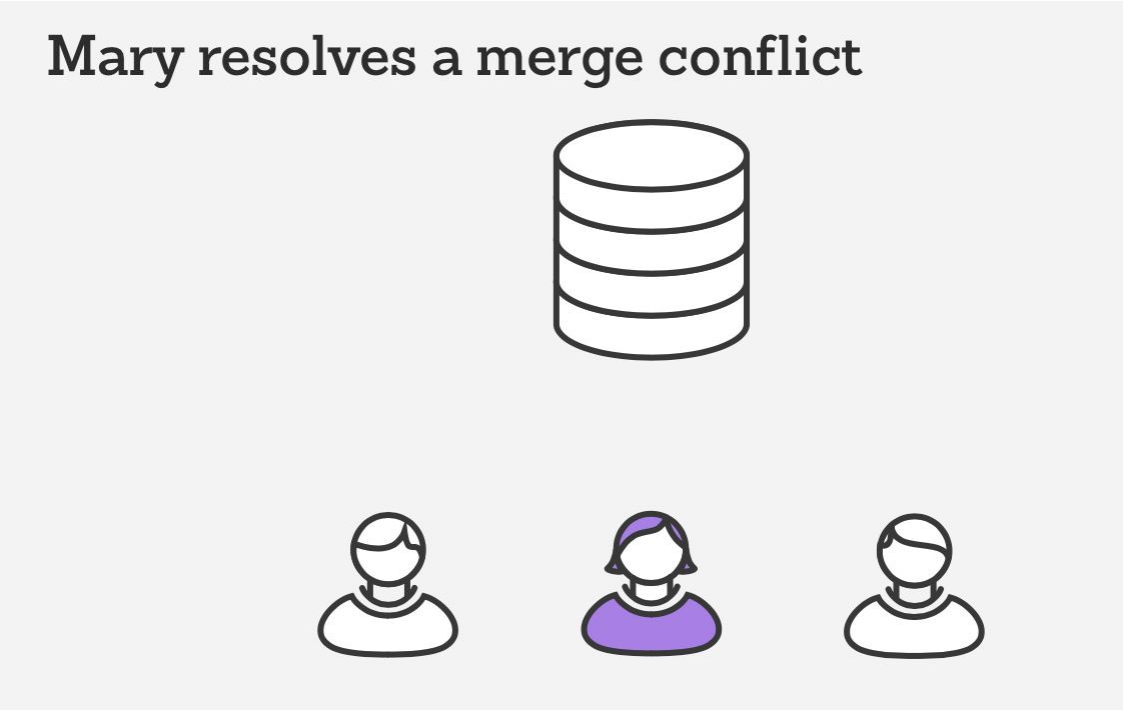


Example

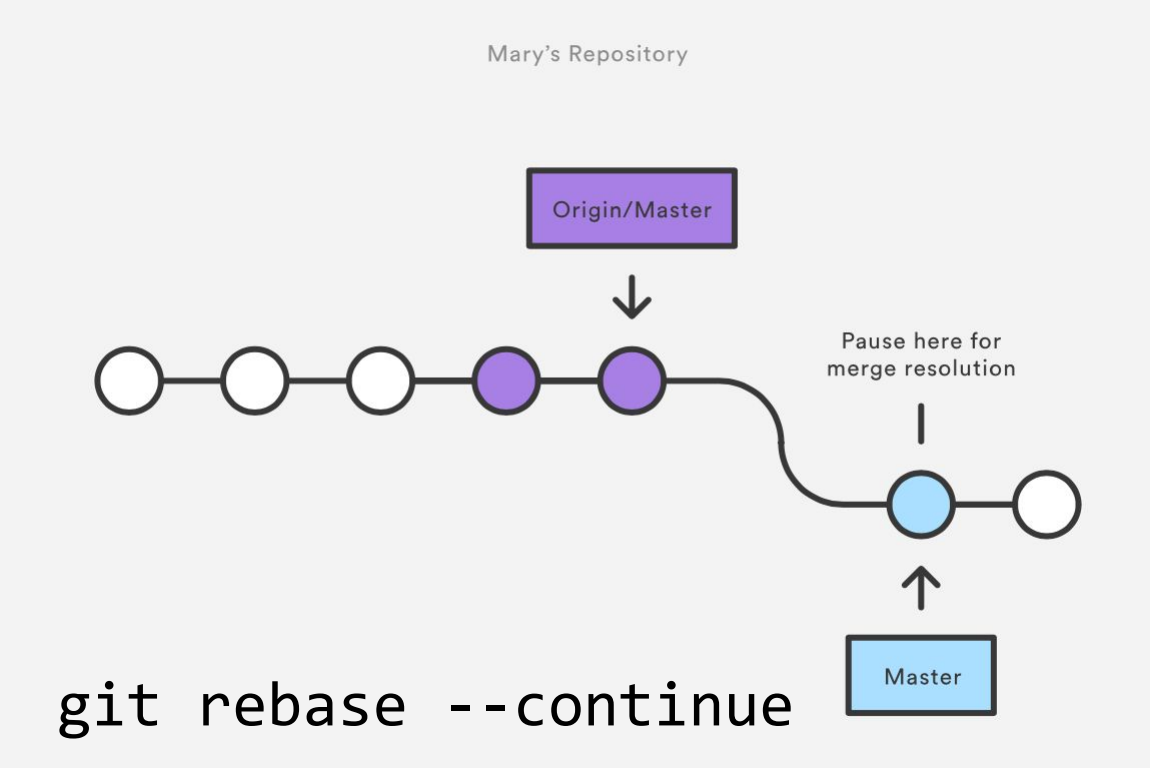

Mary rebases on top of John's commit(s)
git pull --rebase origin main


Mary's Repository
Origin/Master
Master
Master

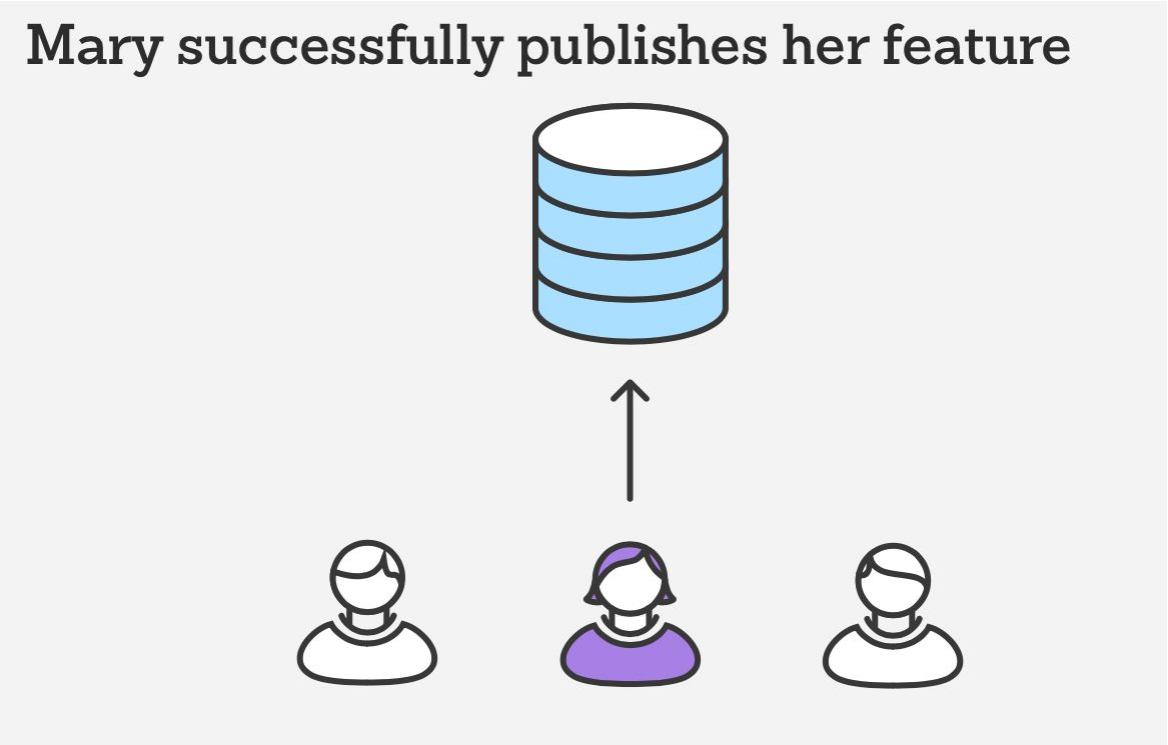
Example



Example


Mary's Repository
Origin/Master
Pause here for
merge resolution
Master
git rebase --continue

Example



2. Git Feature Branch Workflow

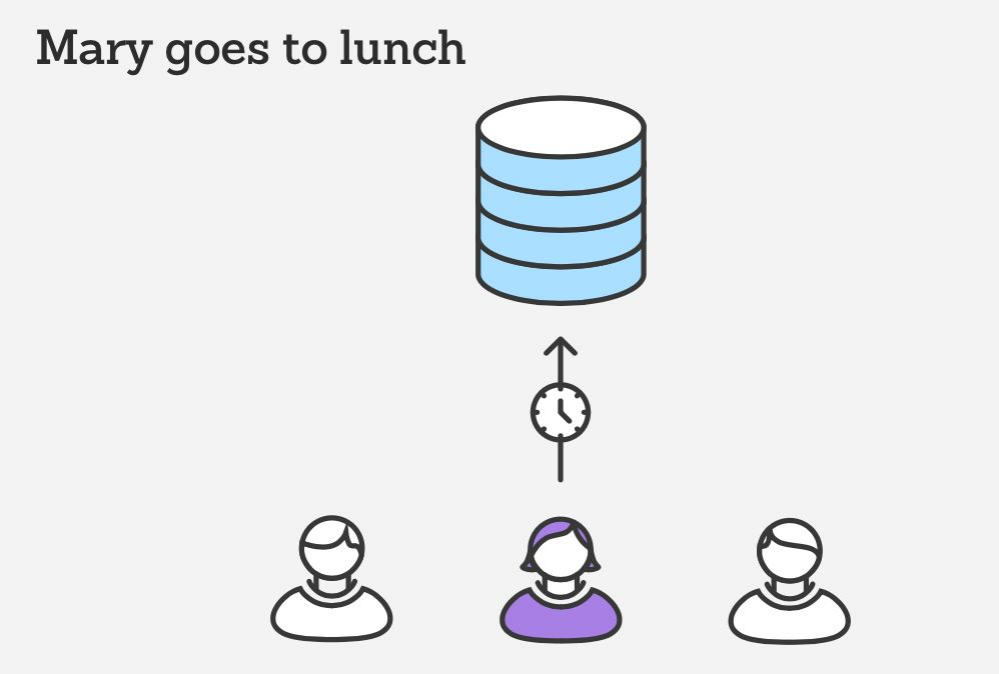
- *All* feature development should take place in a dedicated branch instead of the master branch

- Multiple developers can work on a particular feature without disturbing the main codebase
  - main branch will never contain broken code (enables CI)
  - Enables pull requests (code review)

Example

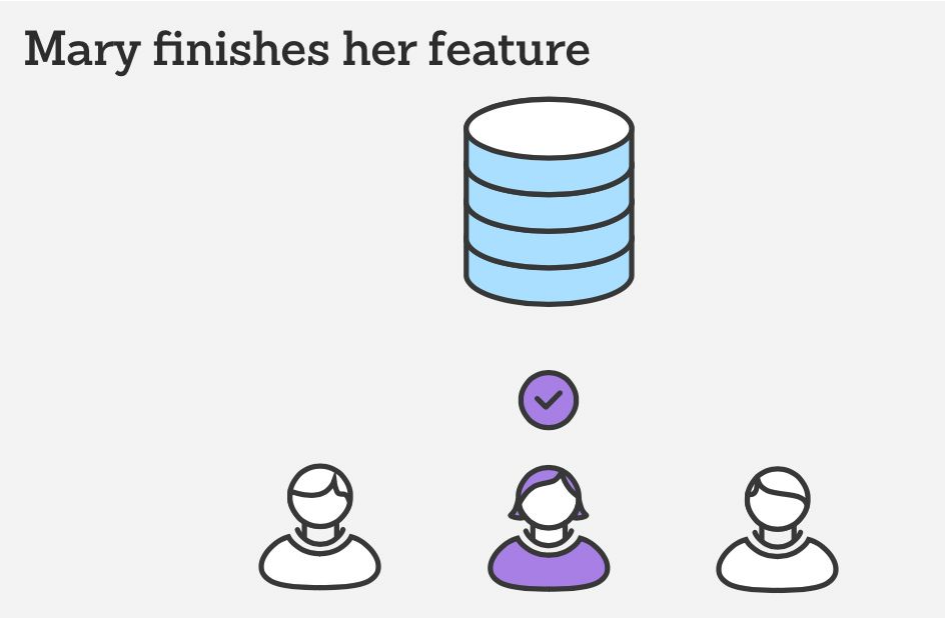

```
git checkout -b marys-feature main
git status
git add <some-file>
git commit
```

Example



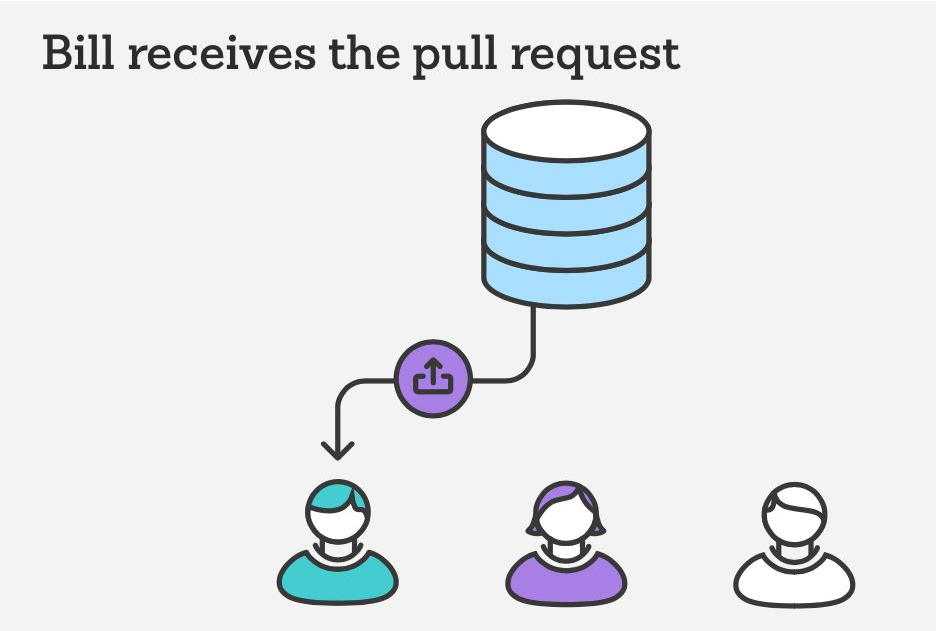

```
git push -u origin marys-feature
```

Example

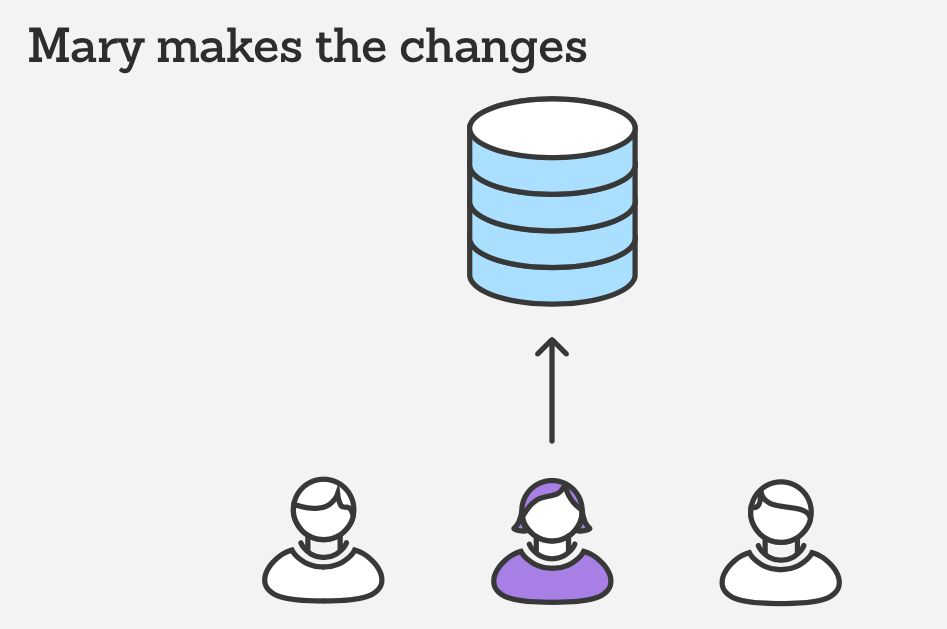



```
git push
```

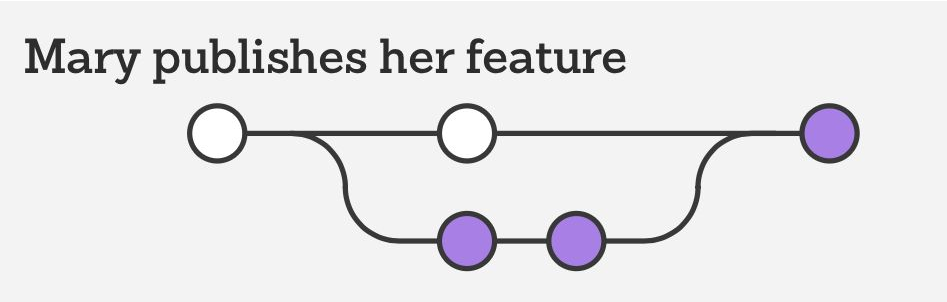
Example



Example

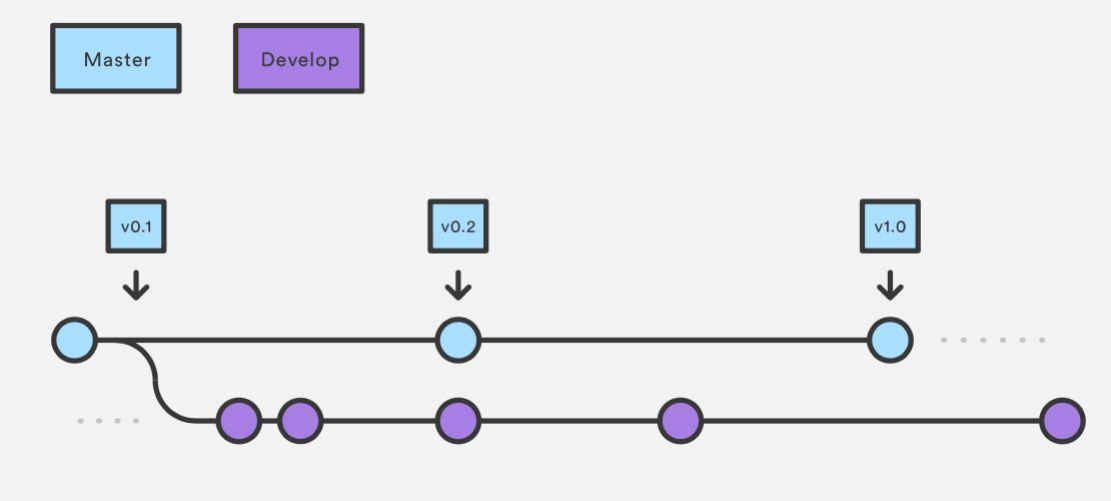


Example - Merge pull request



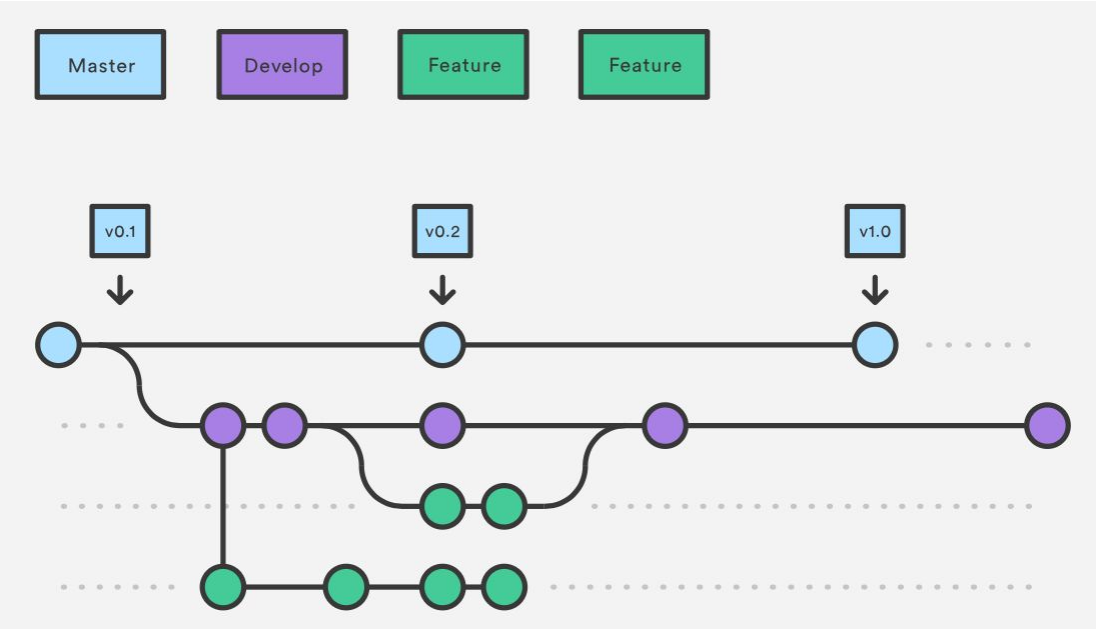

```
git checkout main
git pull
git pull origin marys-feature
git push
```

3. Gitflow Workflow

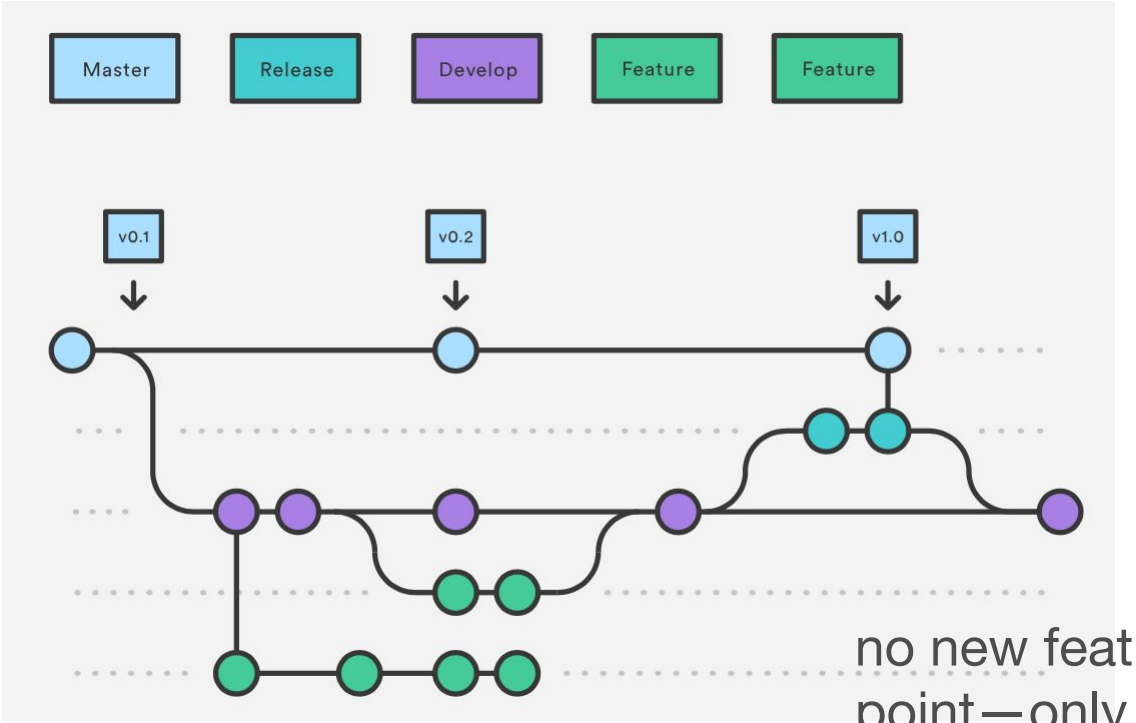


- Strict branching model designed around the project release
  - Suitable for projects that have a scheduled release cycle
- Branches have specific roles and interactions
- Uses two branches
  - main stores the official release history; tag all commits in the main branch with a version number
  - develop serves as an integration branch for features

GitFlow feature branches (from develop)



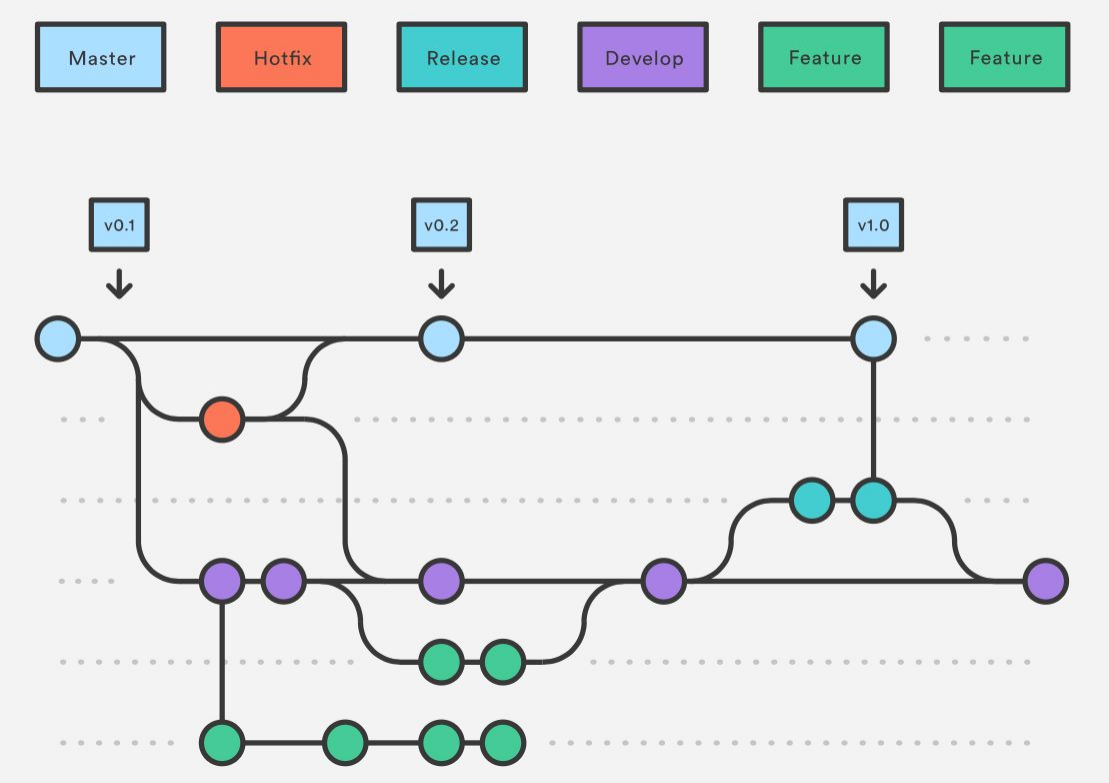
GitFlow release branches (eventually into master)



no new features after this point—only bug fixes, docs, and other release tasks

GitFlow hotfix branches

used to quickly patch production releases



Aside: Semantic Versioning

Semantic Versioning

Given a version number MAJOR.MINOR.PATCH, increment the:

1. MAJOR version when you make incompatible API changes,
2. MINOR version when you add functionality in a backwards compatible manner, and
3. PATCH version when you make backwards compatible bug fixes.

Code status	Stage	Rule	Example version
First release	New product	Start with 1.0.0	1.0.0
Backward compatible bug fixes	Patch release	Increment the third digit	1.0.1
Backward compatible new features	Minor release	Increment the middle digit and reset last digit to zero	1.1.0
Changes that break backward compatibility	Major release	Increment the first digit and reset middle and last digits to zero	2.0.0

https://docs.npmjs.com/about-semantic-versioning

Summary

- Version control has many advantages
  - History, traceability, versioning
  - Collaborative and parallel development
- Collaboration with branches
  - Different workflows
- From local to central to distributed version control

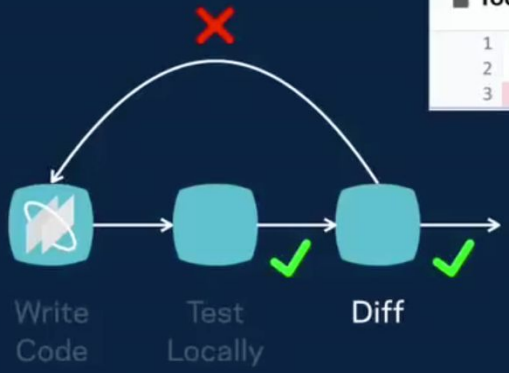
DEVELOPMENT AT SCALE

Releasing at scale in industry

- Facebook: https://atscaleconference.com/videos/rapid-release-at-massive-scale/
- Google: https://www.slideshare.net/JohnMicco1/2016-0425-continuous-integration-at-google-scal
  https://testing.googleblog.com/2011/06/testing-at-speed-and-scale-of-google.html
- Why Google Stores Billions of Lines of Code in a Single Repository: https://www.youtube.com/watch?v=W71BTkUbdqE
- F8 2015 - Big Code: Developer Infrastructure at Facebook's Scale: https://www.youtube.com/watch?v=X0VH78ye4yY

Pre-2017 release management model at Facebook

Diff lifecycle: local testing



Tools/xctool/xctool/xctool/Version.m

View Options

```
1  #import "Version.h"                                  1  #import "Version.h"
2                                                       2
3  NSString * const XCToolVersionString = @"0.2.1";    3  NSString * const XCToolVersionString = @"0.2.2";
```

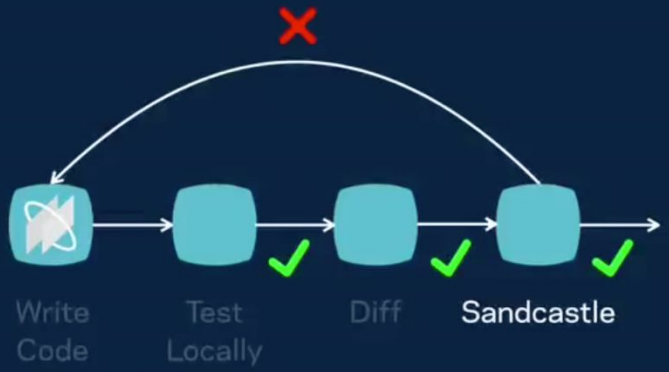
```
PASS ExampleTest (0.050s)
.

OK (1 test, 4 assertions)

OK
(1 tests, 4 assertions, 0 incomplete, 0 failures)
```

Test and lint locally

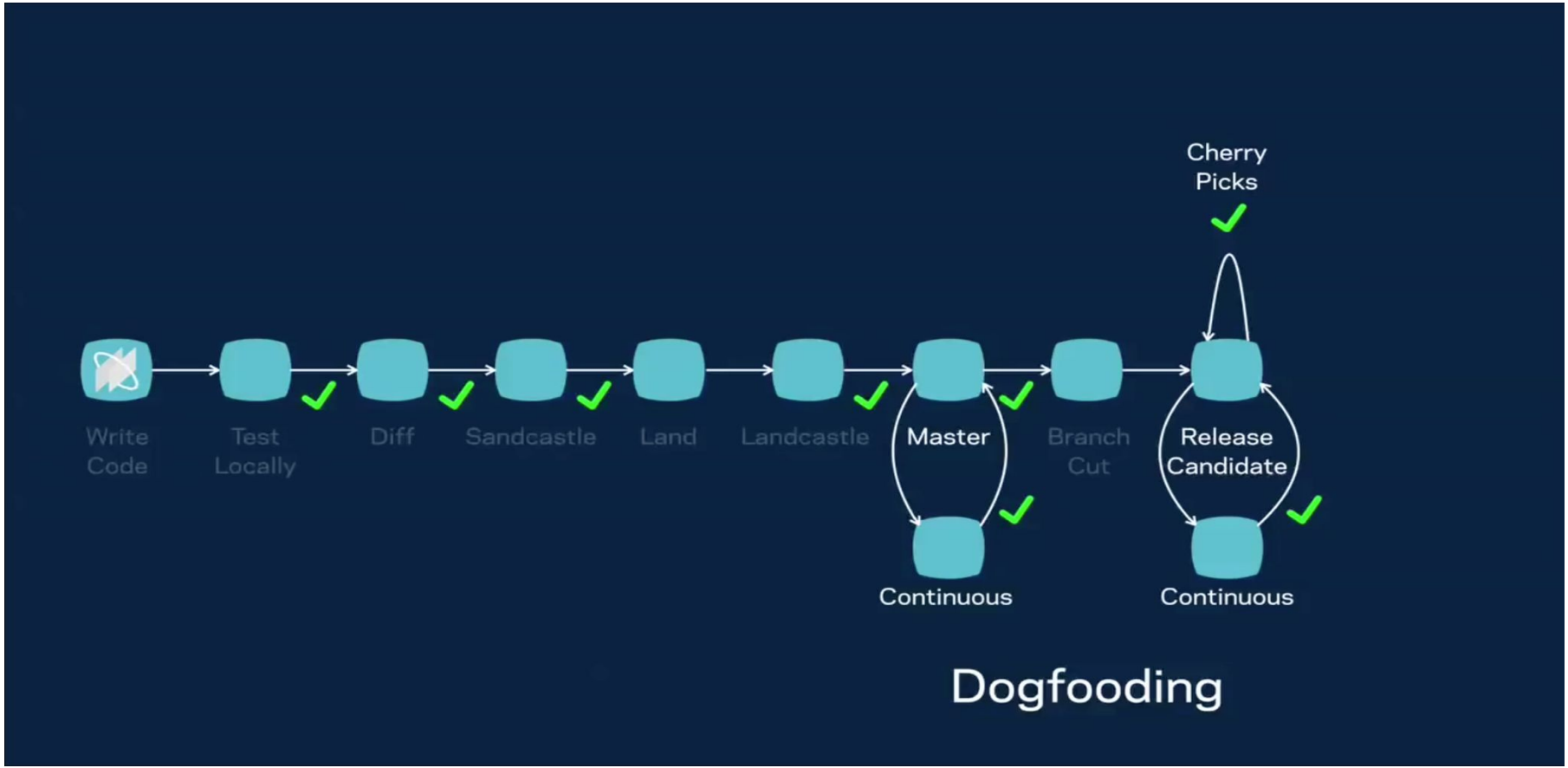
Diff lifecycle: CI testing (data center)



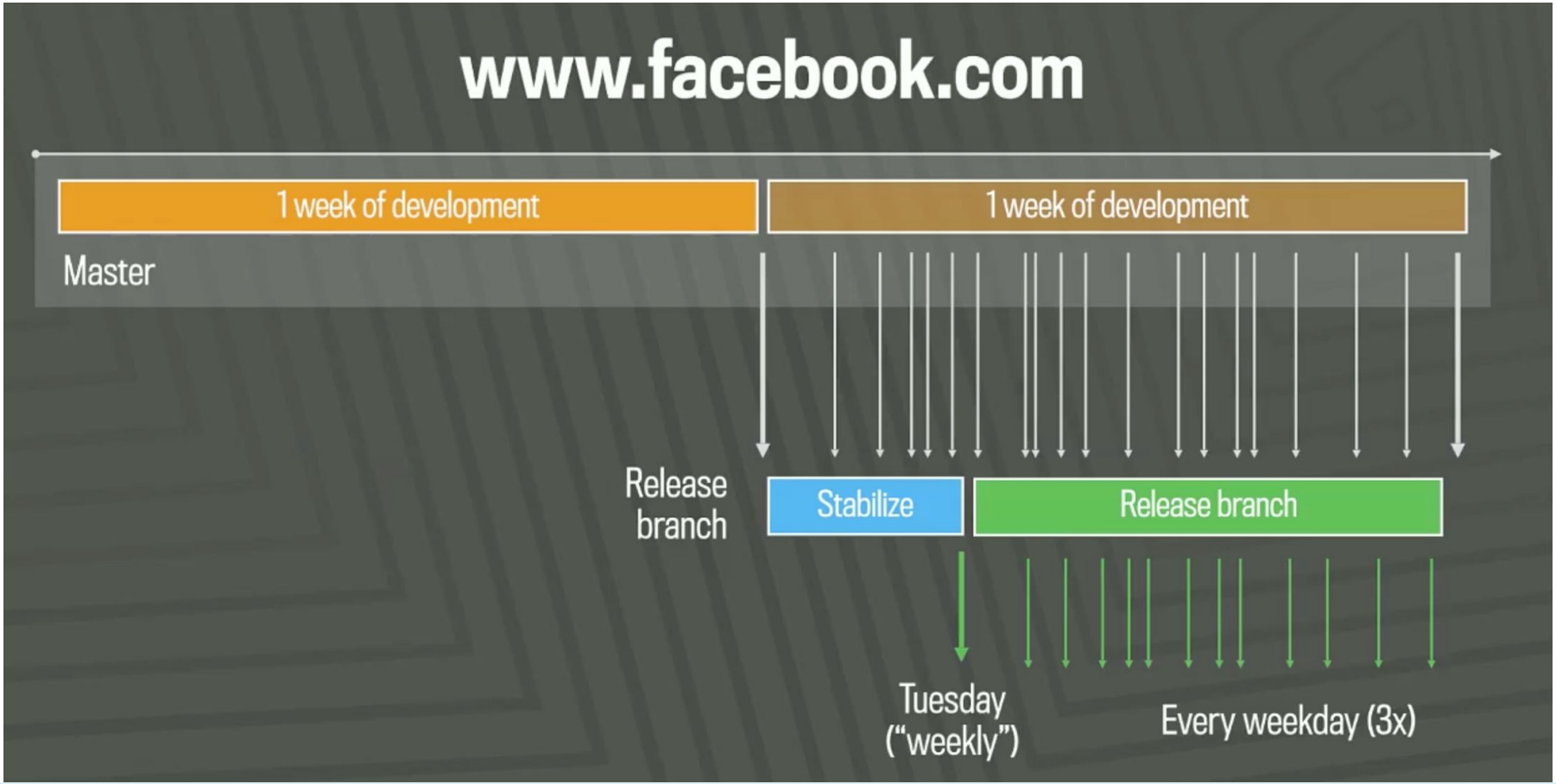
	Facebook	Messenger	Groups	...
arm	✓	✓	✓	✓
x86	✓	✓	✓	✓
...	✓	✓	✓	✓

App and Build Configuration Matrix

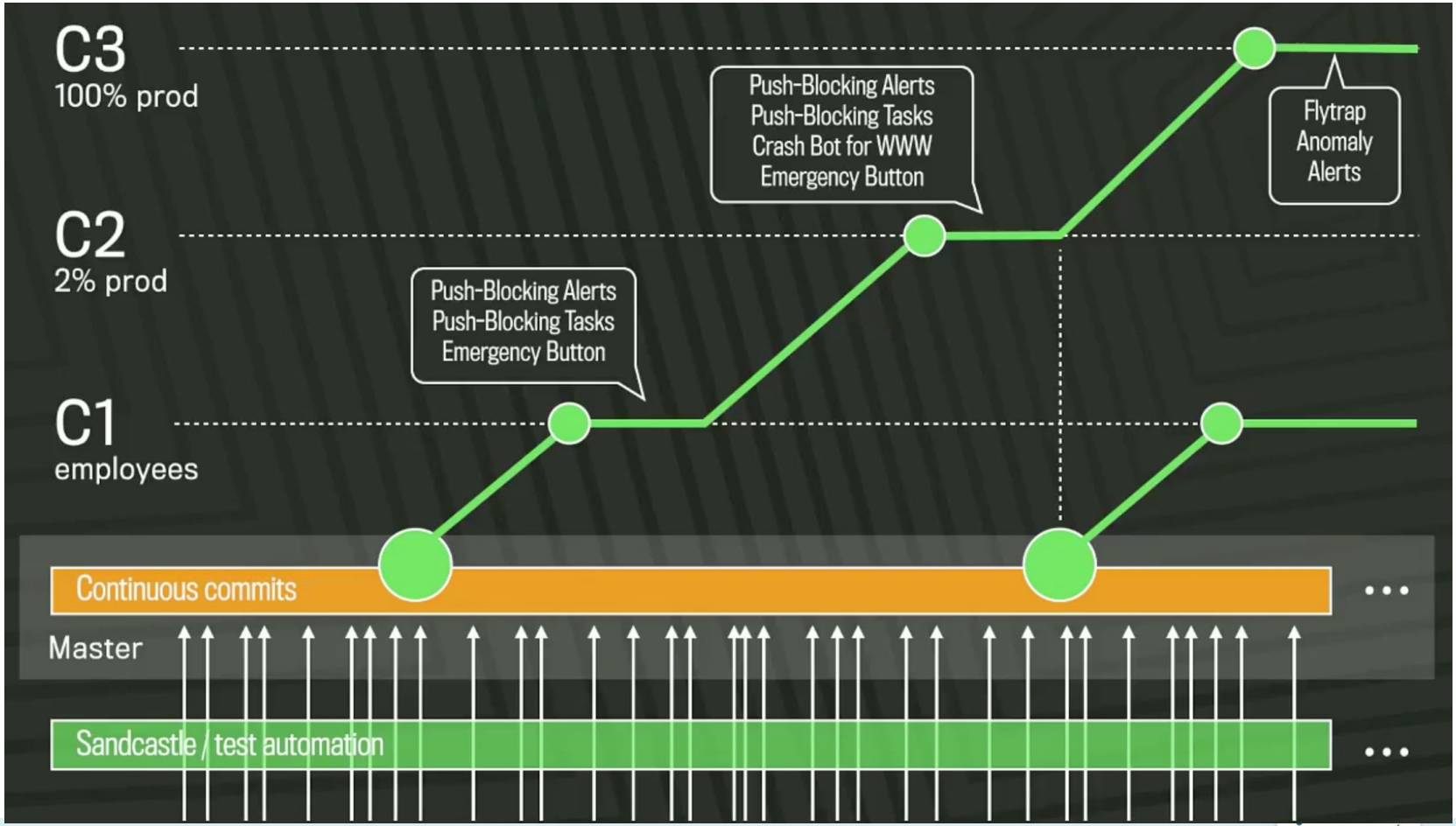
Diff lifecycle: diff ends up on main branch



Release every two weeks



Quasi-continuous push from master (1,000+ devs, 1,000 diffs/day); 10 pushes/day



https://samritchie.wordpress.com/2013/10/16/build-server-traffic-lights/

https://www.softwire.com/blog/2013/09/26/continuous-integration-traffic-lights-revamp/index.html

Status

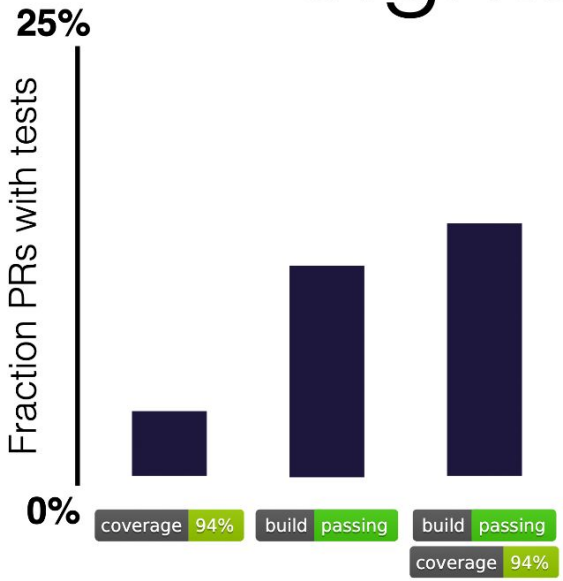
Build Pipeline

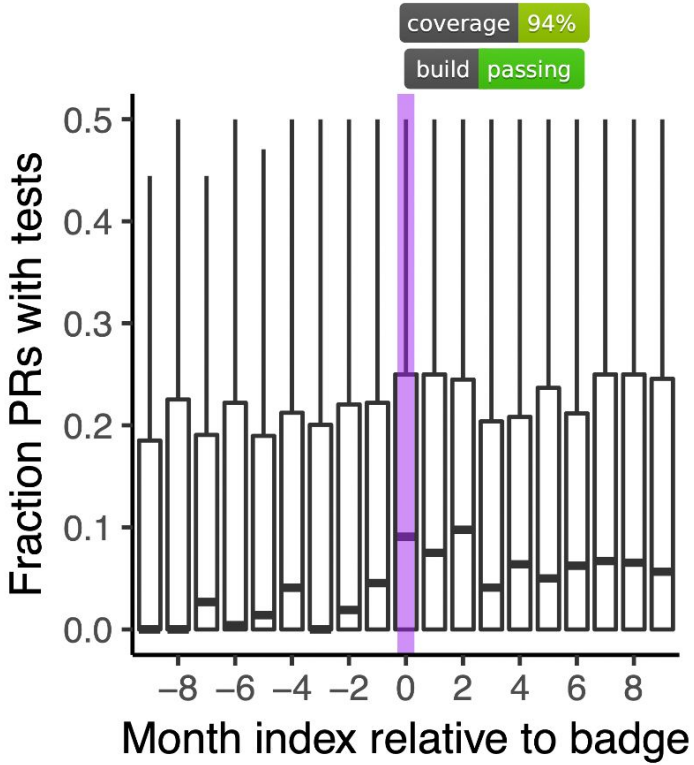
Release Pipeline

Dev	Test	Prod
		
		

https://blog.devops4me.com/status-badges-in-azure-devops-pipelines/

Signals of PR quality





Result: Build status+code coverage badges indicate *more tests in PRs*

STRUDEL
Carnegie Mellon University

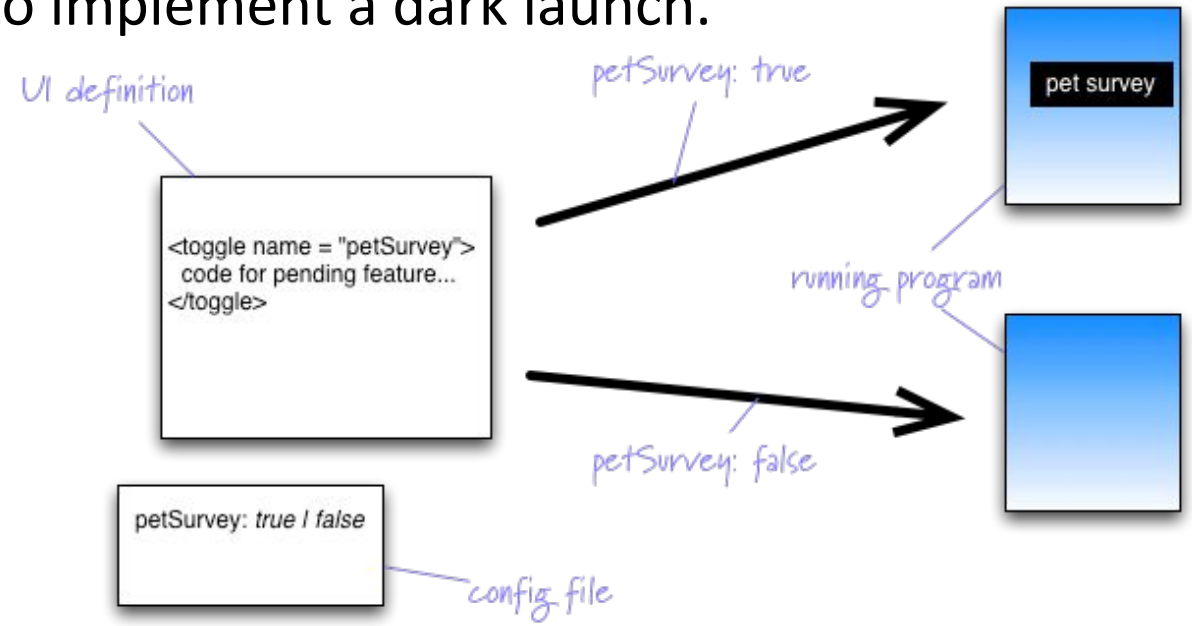
Aside: Key idea – fast to deploy, slow to release

Dark launches at Instagram

- **Early**: Integrate as soon as possible. Find bugs early. Code can run in production about 6 months before being publicly announced (“dark launch”).
- **Often**: Reduce friction. Try things out. See what works. Push small changes just to gather metrics, feasibility testing. Large changes just slow down the team. Do dark launches, to see what performance is in production, can scale up and down. *"Shadow infrastructure" is too expensive, just do in production.*
- **Incremental**: Deploy in increments. Contain risk. Pinpoint issues.

Aside: Feature Flags

Typical way to implement a dark launch.



http://swreflections.blogspot.com/2014/08/feature-toggles-are-one-of-worst-kinds.html

http://martinfowler.com/bliki/FeatureToggle.html

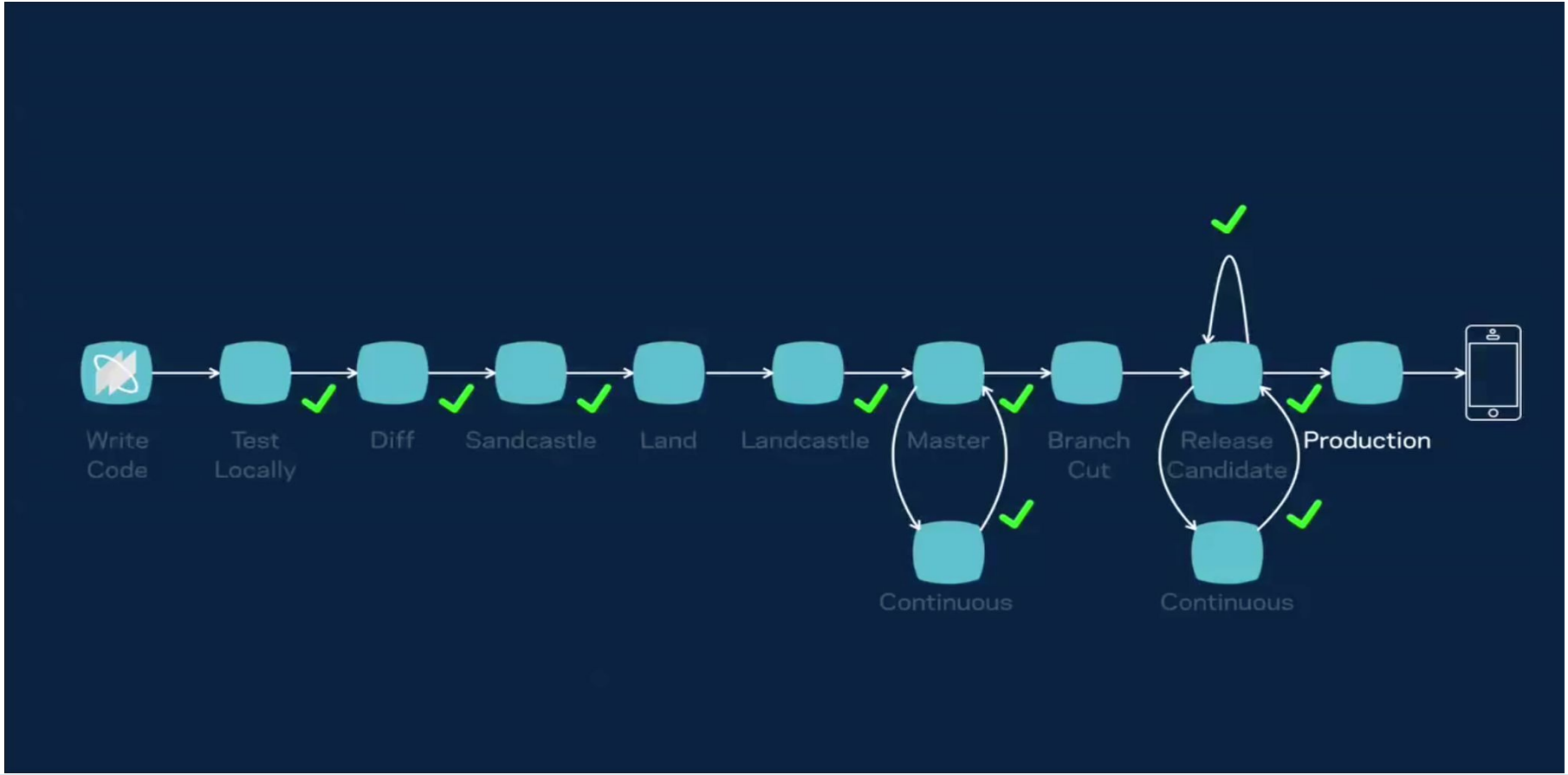
Issues with feature flags

Feature flags are “technical debt”

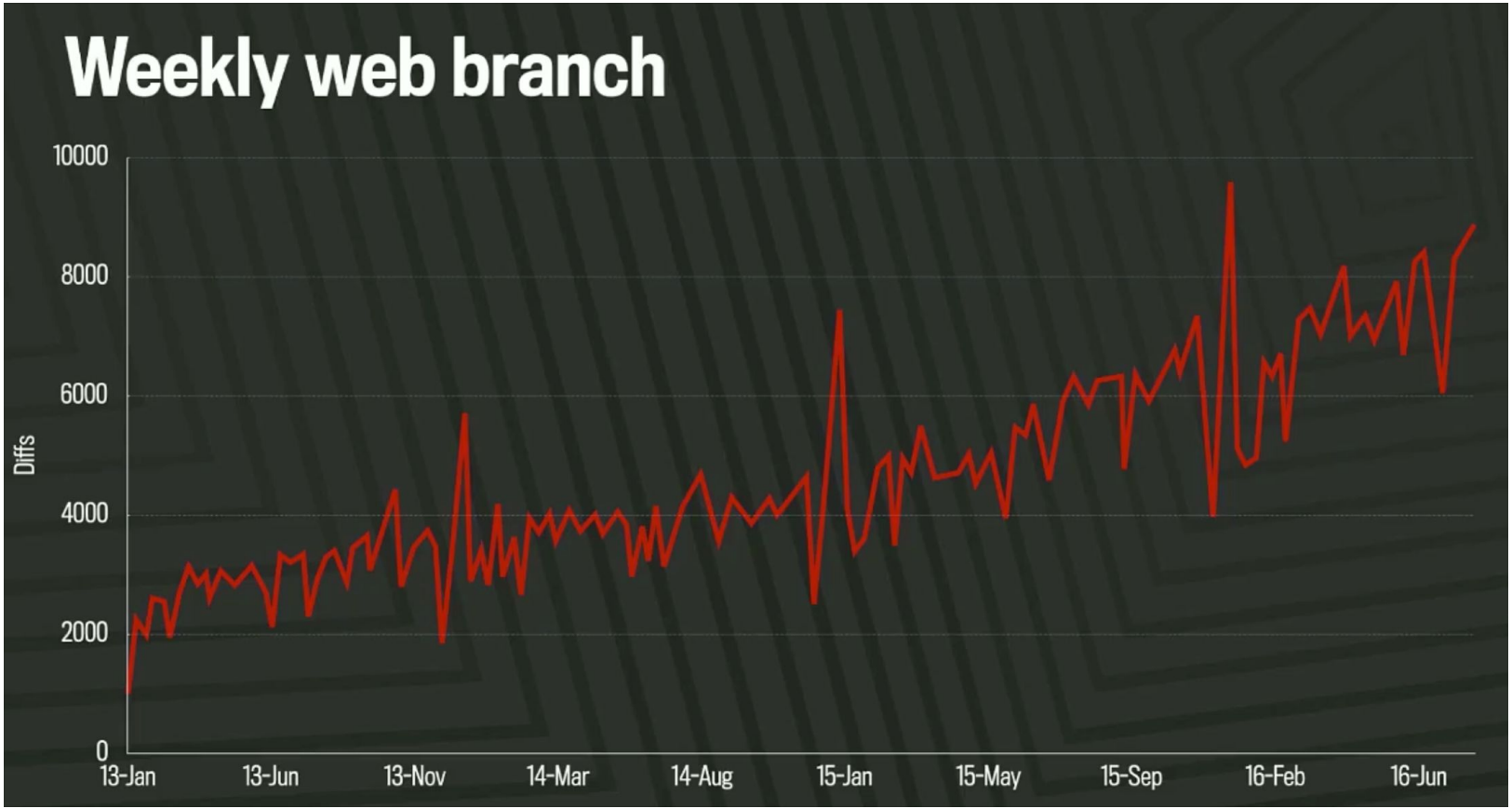
Example: $400 million financial services company went bankrupt in 45 minutes.

http://dougseven.com/2014/04/17/knightmare-a-devops-cautionary-tale/

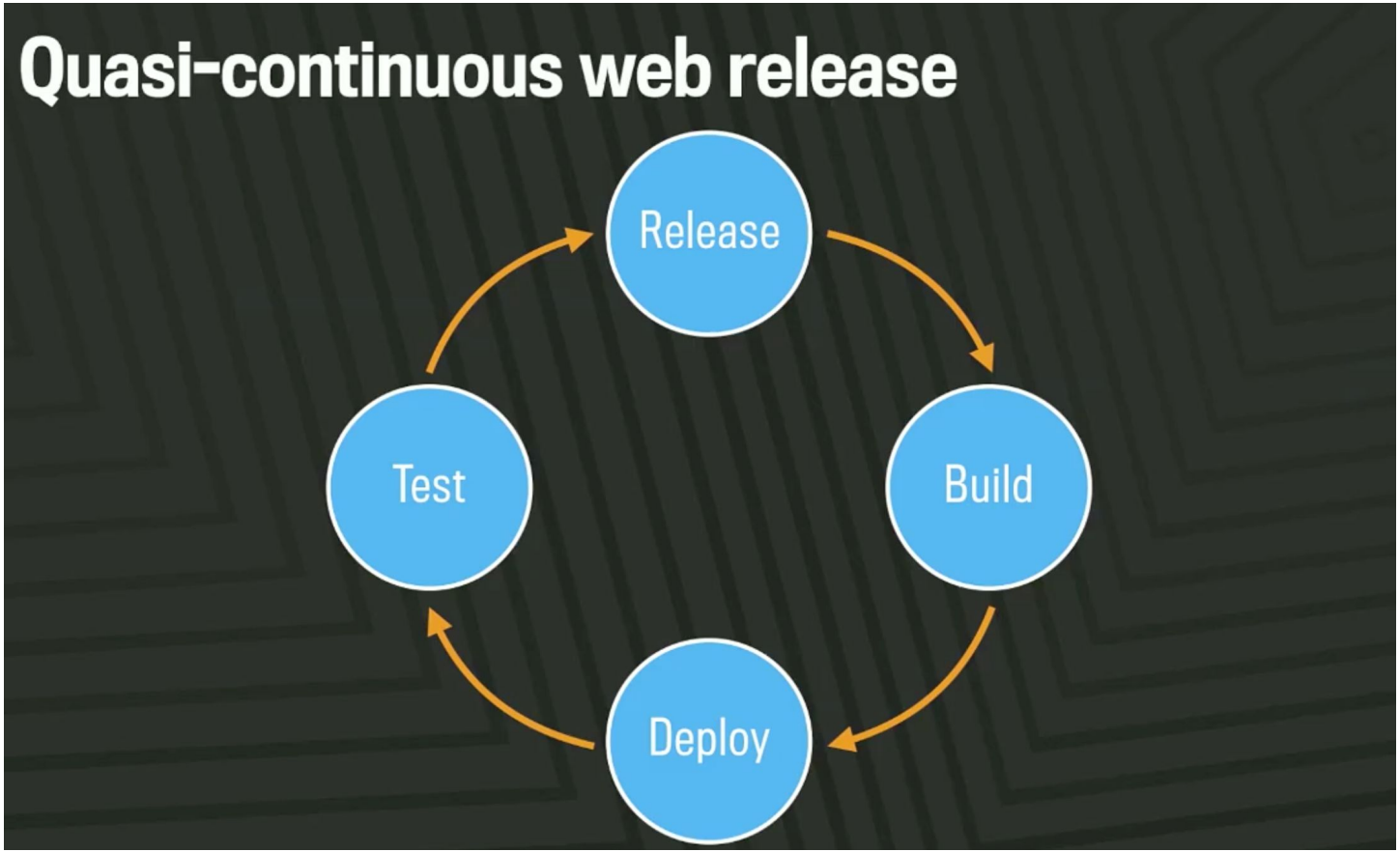
Diff lifecycle: in production



What's in a weekly branch cut? (The limits of branches)



Post-2017 release management model at Facebook



Google: similar story. YUGE code base

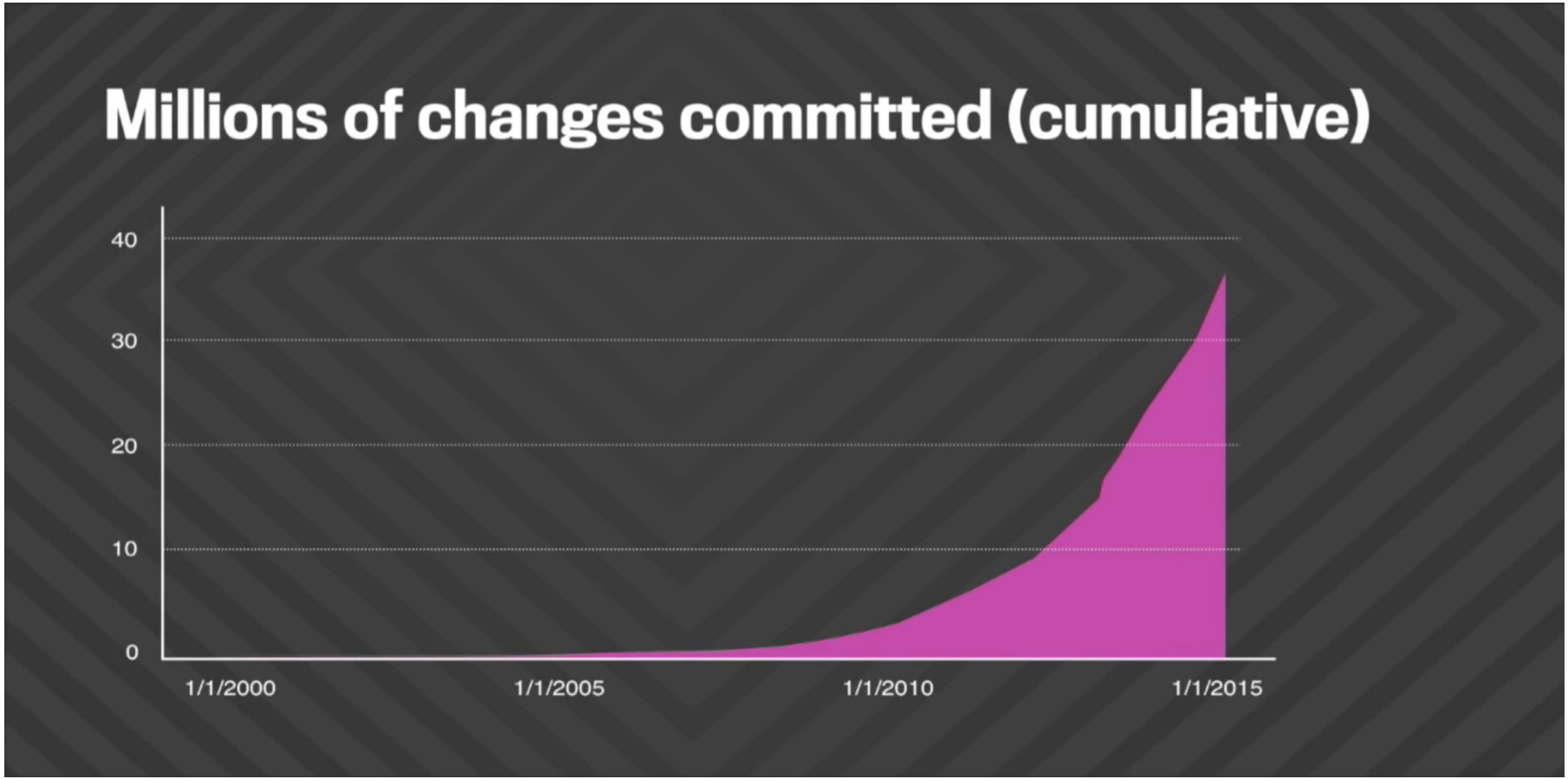
Google repository statistics

As of Jan 2015

Total number of files*	1 billion
Number of source files	9 million
Lines of code	2 billion
Depth of history	35 million commits
Size of content	86 terabytes
Commits per workday	45 thousand

*The total number of files includes source files copied into release branches, files that are deleted at the latest revision, configuration files, documentation, and supporting data files.

Exponential growth



Google Speed and Scale

- >30,000 developers in 40+ offices
- 13,000+ projects under active development
- 30k submissions per day (1 every 3 seconds)

- All builds from source
- 30+ sustained code changes per minute with 90+ peaks
- 50% of code changes monthly
- 150+ million test cases / day, > 150 years of test / day
- Supports continuous deployment for all Google teams!

Google code base vs Linux kernel code base

Some perspective

Linux kernel

- 15 million lines of code in 40 thousand files (total)

Google repository

- 15 million lines of code in 250 thousand files *changed per week, by humans*
- 2 billion lines of code, in 9 million source files (total)

How do they do it?

1. Lots of (automated) testing

Google workflow

- All code is reviewed before commit (by humans and automated tooling)
- Each directory has a set of owners who must approve the change to their area of the repository
- Tests and automated checks are performed before and after commit
- Auto-rollback of a commit may occur in the case of widespread breakage

2. Lots of automation

Additional tooling support

Critique	Code review
CodeSearch*	Code browsing, exploration, understanding, and archeology
Tricorder**	Static analysis of code surfaced in Critique, CodeSearch
Presubmits	Customizable checks, testing, can block commit
TAP	Comprehensive testing before and after commit, auto-rollback
Rosie	Large-scale change distribution and management

* See "How Developers Search for Code: A Case Study", In European Software Engineering Conference and the ACM SIGSOFT Symposium on the Foundations of Software Engineering, 2015
** See "Tricorder: Building a program analysis ecosystem". In International Conference on Software Engineering (ICSE), 2015

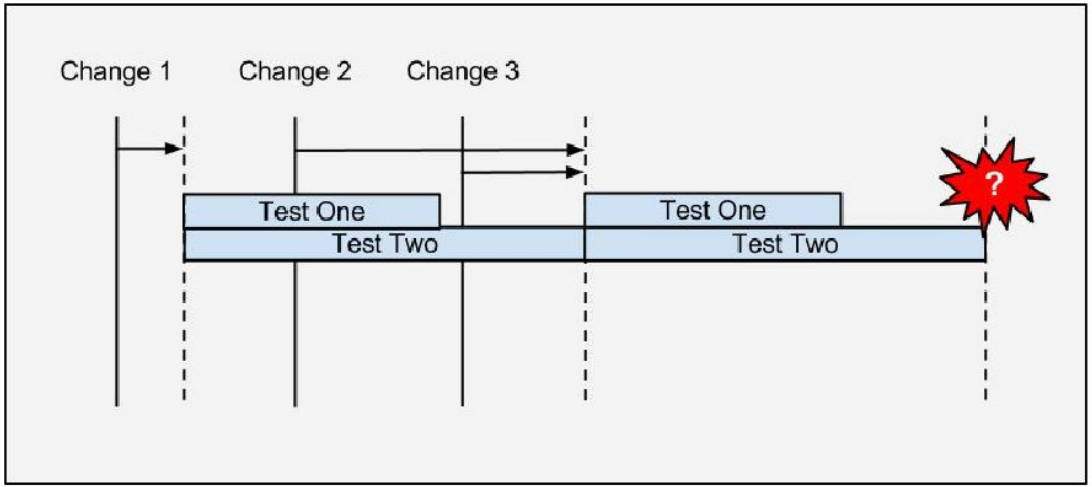
3. Smarter tooling

- Build system
- Version control
- …

3a. Build system

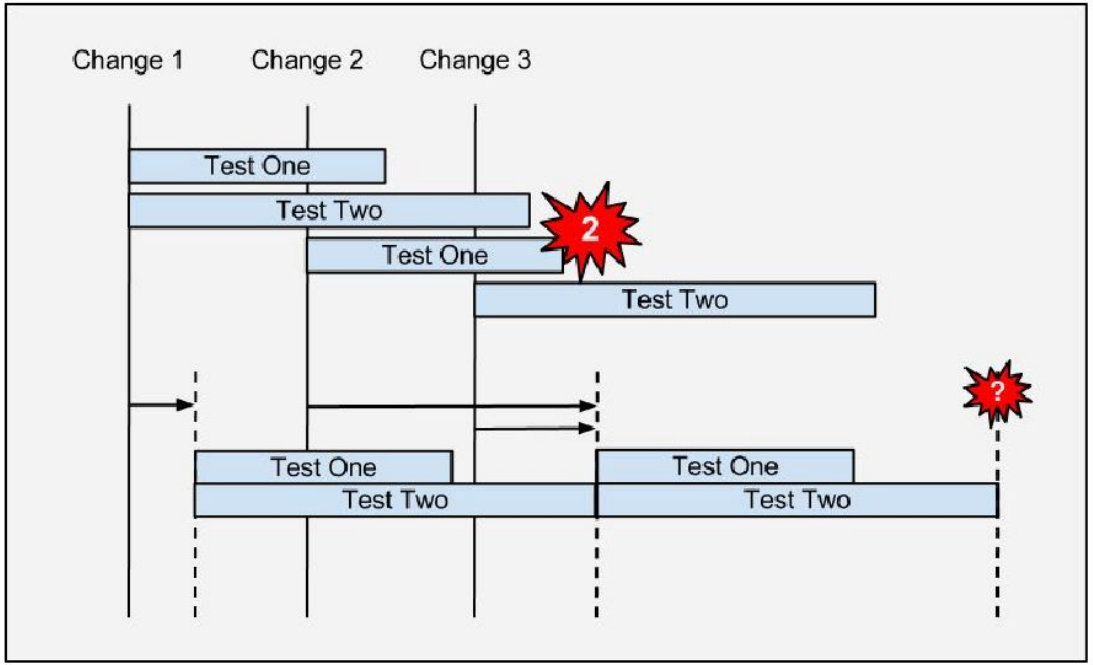
Google Standard Continuous Build System

- Triggers builds in continuous cycle
- Cycle time = longest build + test cycle
- Tests many changes together
- Which change broke the build?

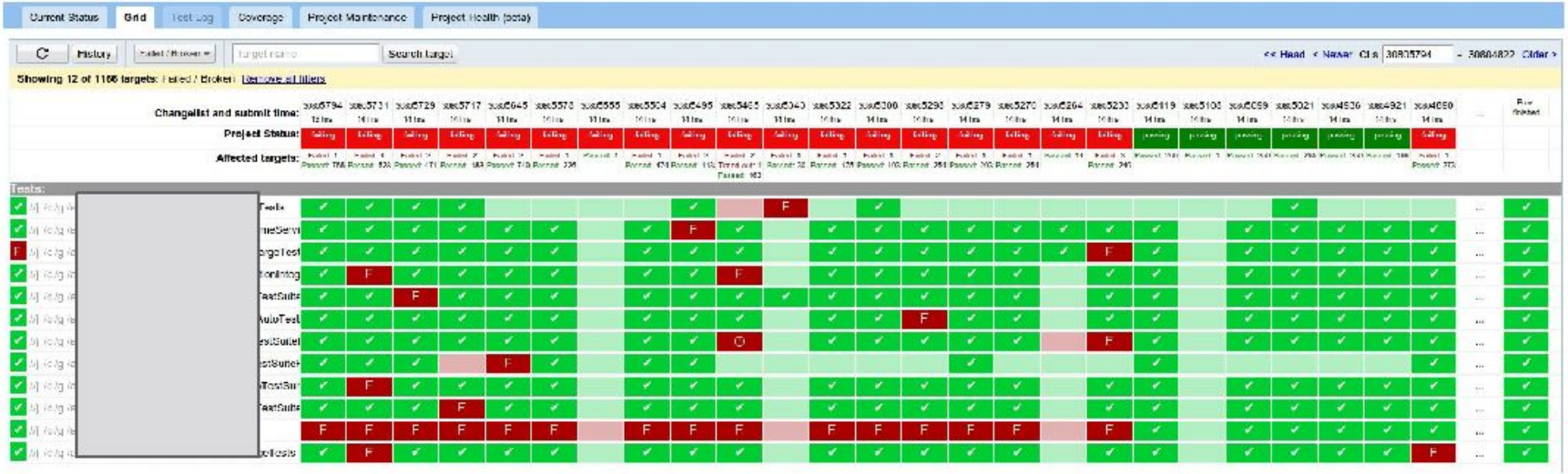


Google Continuous Build System

- Triggers tests on every change
- Uses fine-grained dependencies
- Change 2 broke test 1



Google Continuous Integration Display



Google Benefits

- Identifies failures sooner
- Identifies culprit change precisely
  - Avoids divide-and-conquer and tribal knowledge
- Lower compute costs using fine grained dependencies
- Keeps the build green by reducing time to fix breaks
- Accepted enthusiastically by product teams
- Enables teams to ship with fast iteration times
  - Supports submit-to-production times of less than 36 hours for some projects

Google Costs

- Requires enormous investment in compute resources (it helps to be at Google) grows in proportion to:
  - Submission rate
  - Average build + test time
  - Variants (debug, opt, valgrind, etc.)
  - Increasing dependencies on core libraries
  - Branches
- Requires updating dependencies on each change
  - Takes time to update - delays start of testing

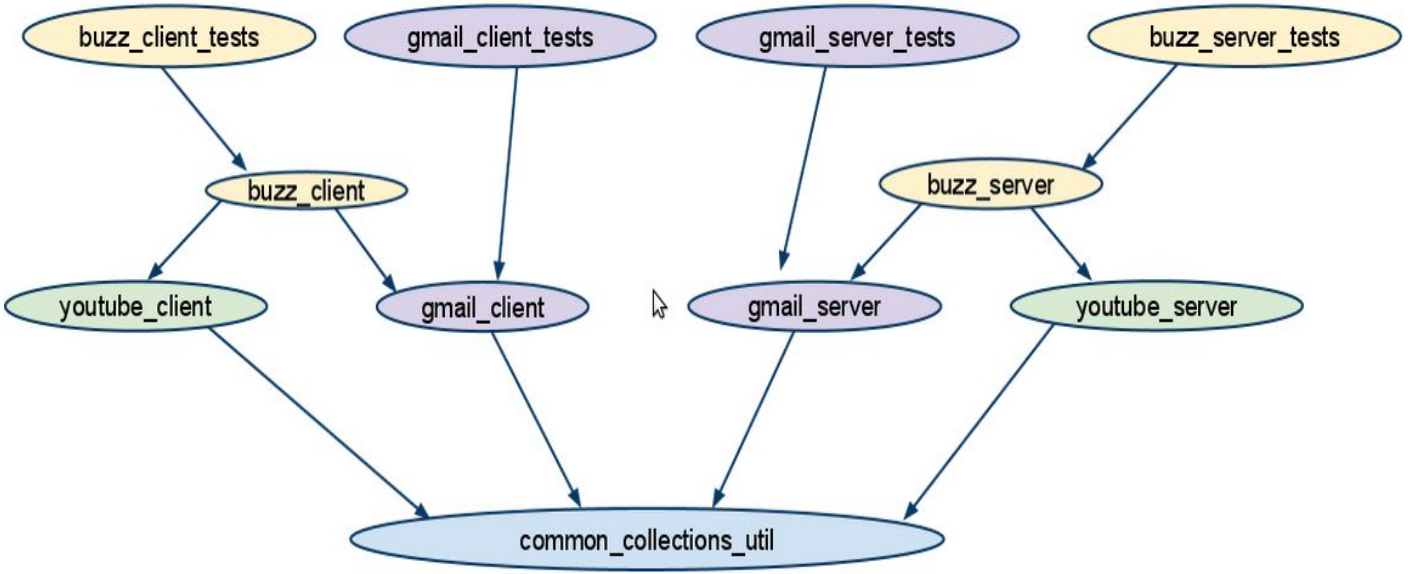
Which tests to run?

GMAIL

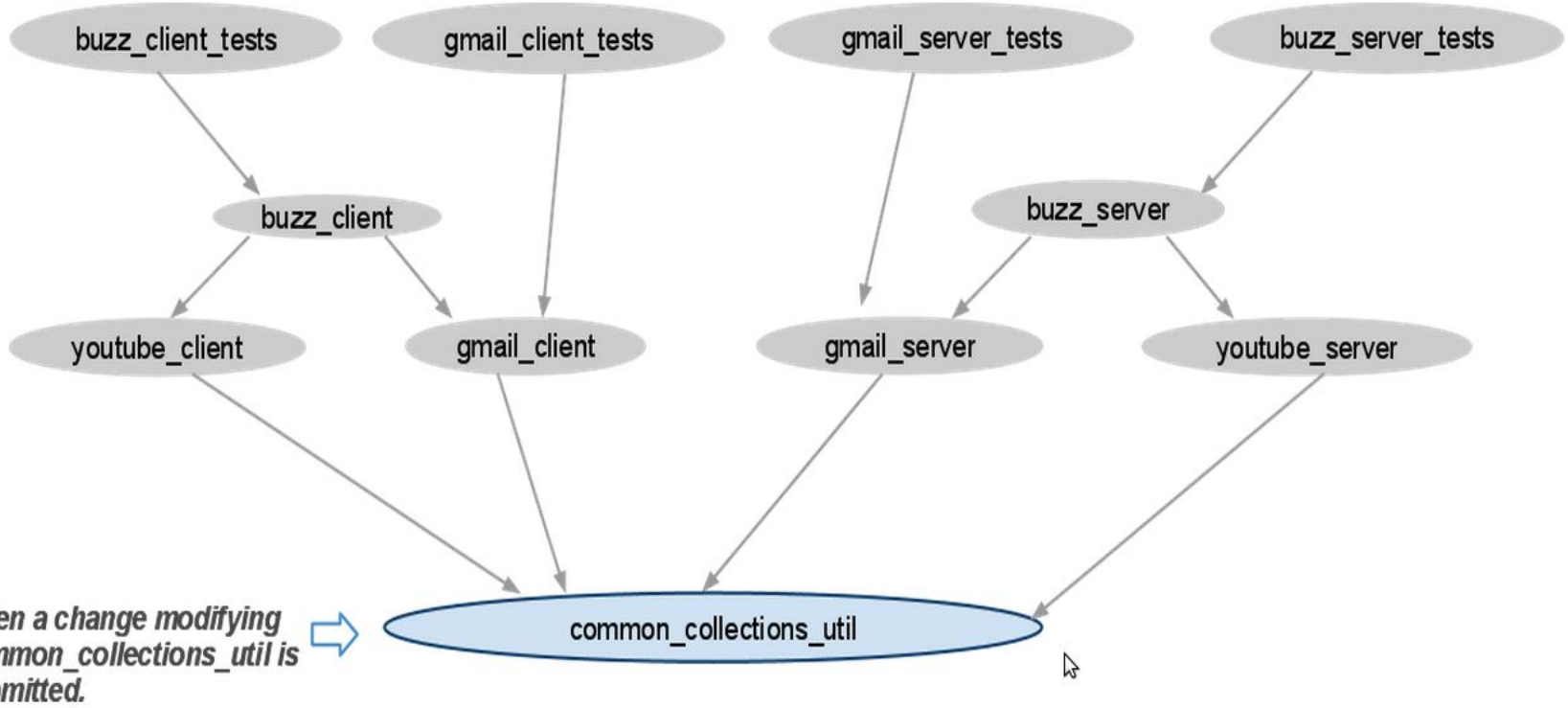
Test Target:
name: //depot/gmail_client_tests
name: //depot/gmail_server_tests

BUZZ

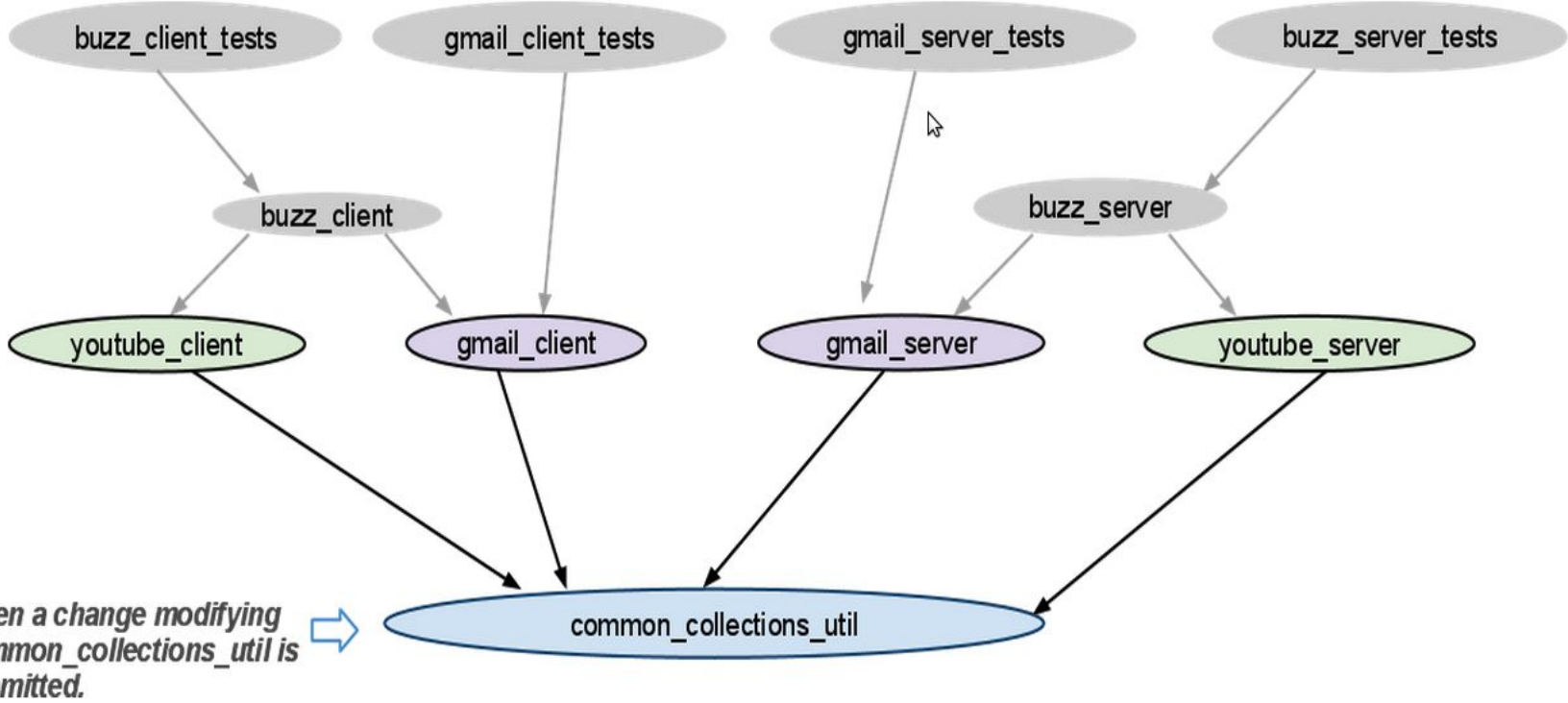
Test targets:
name: //depot/buzz_server_tests
name: //depot/buzz_client_tests



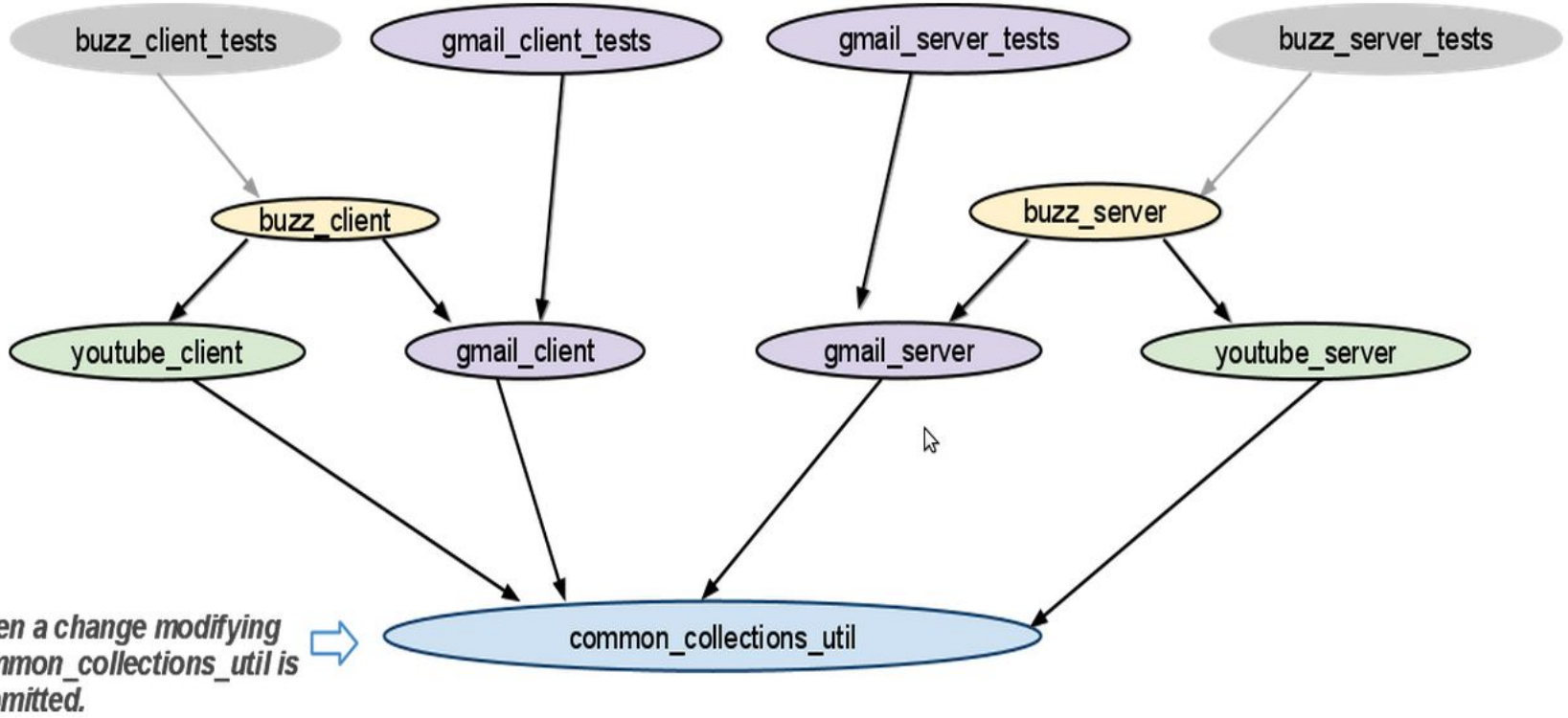
Scenario 1: a change modifies common_collections_util



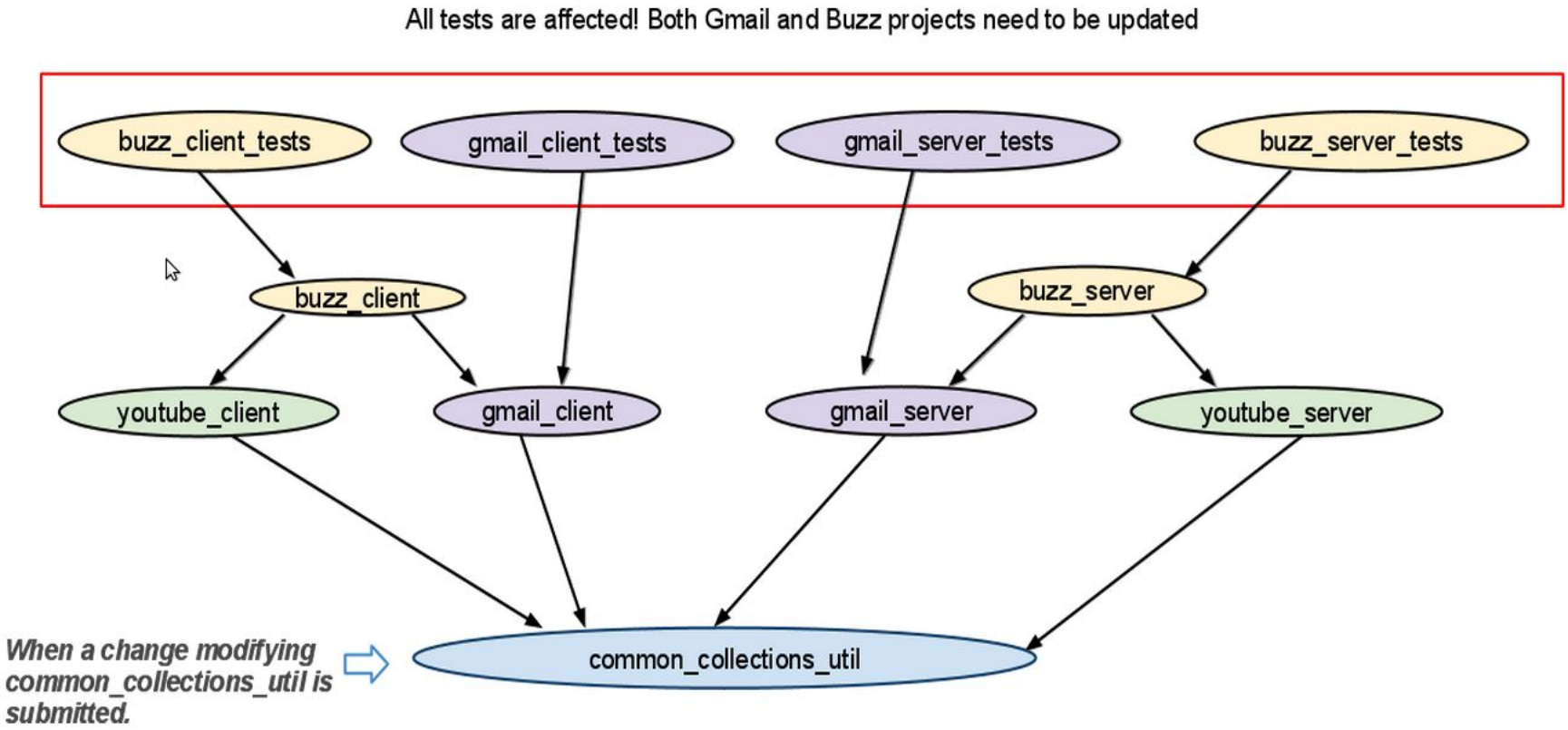
Scenario 1: a change modifies common_collections_util



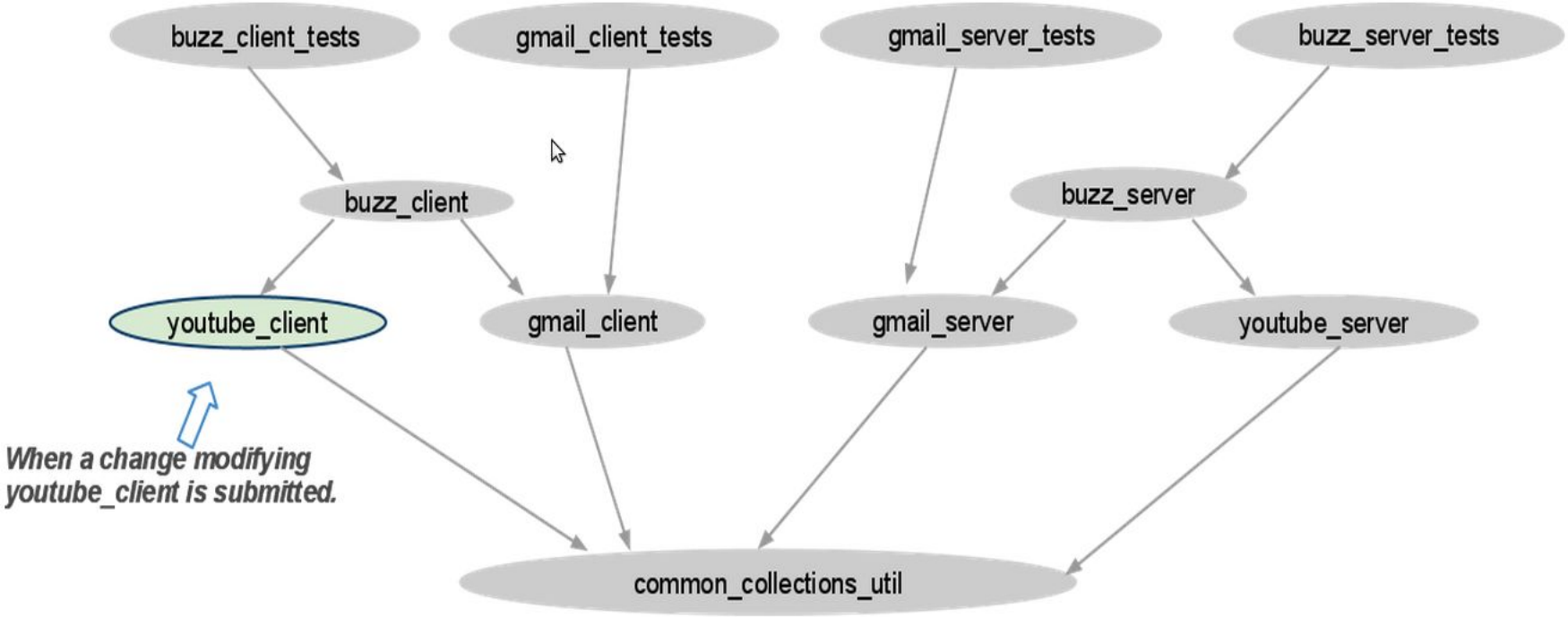
Scenario 1: a change modifies common_collections_util



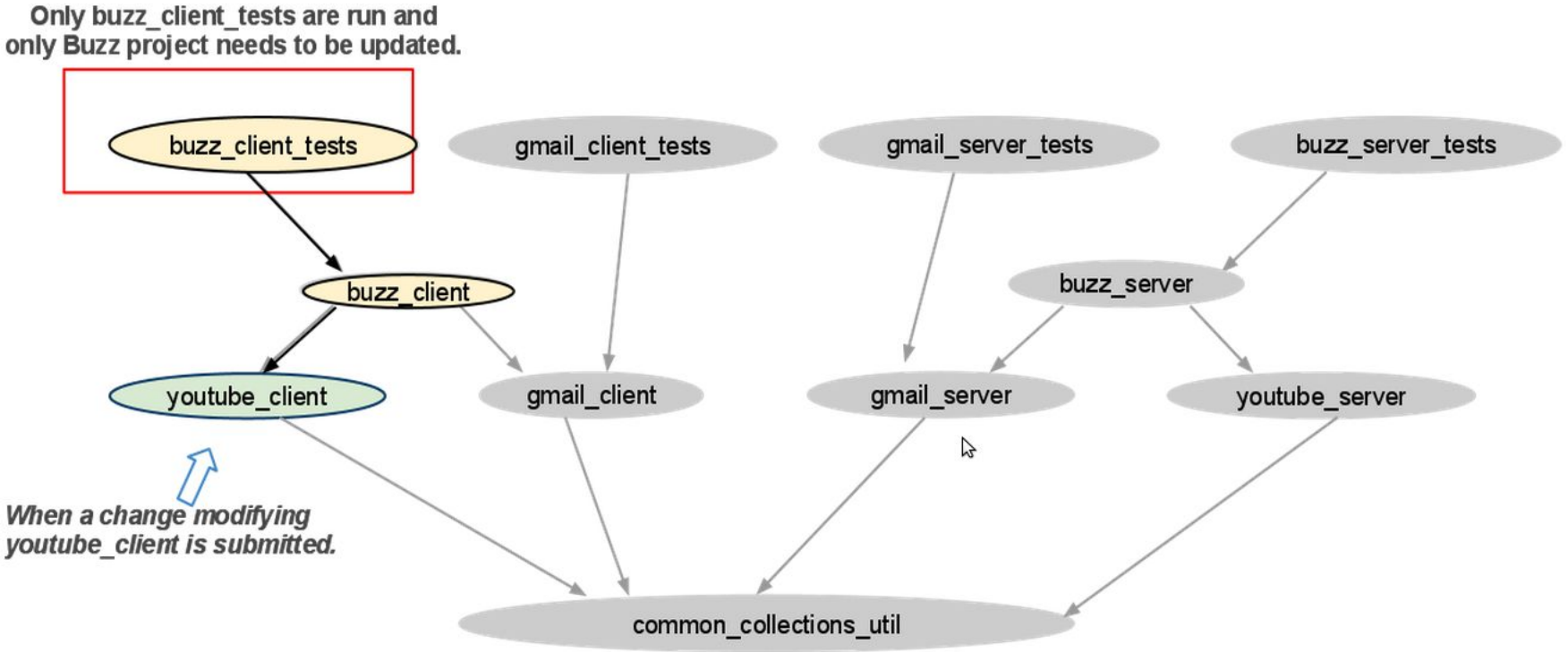
Scenario 1: a change modifies common_collections_util



Scenario 2: a change modifies the youtube_client

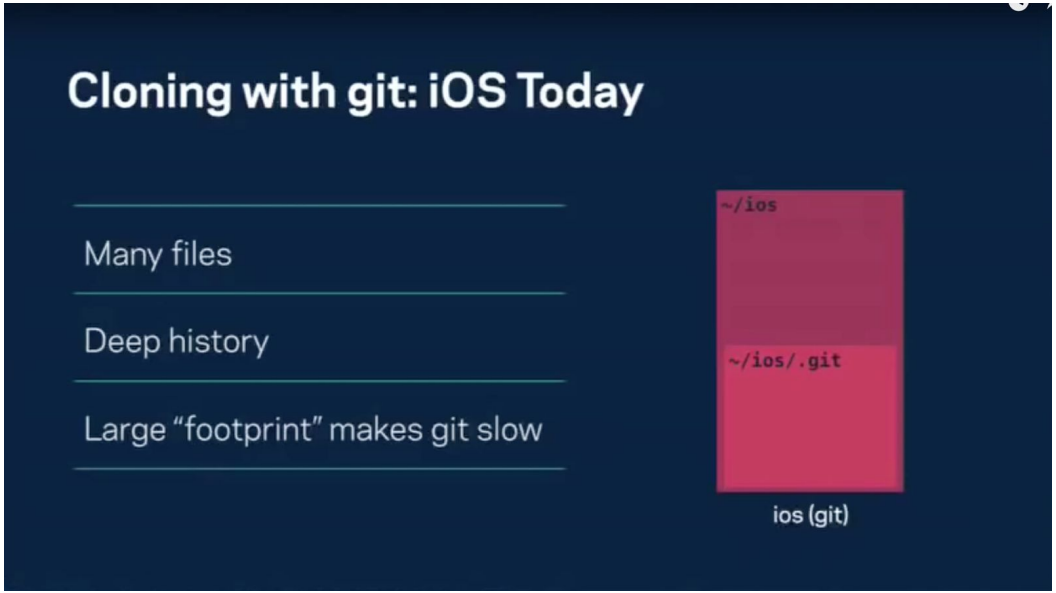


Scenario 2: a change modifies the youtube_client



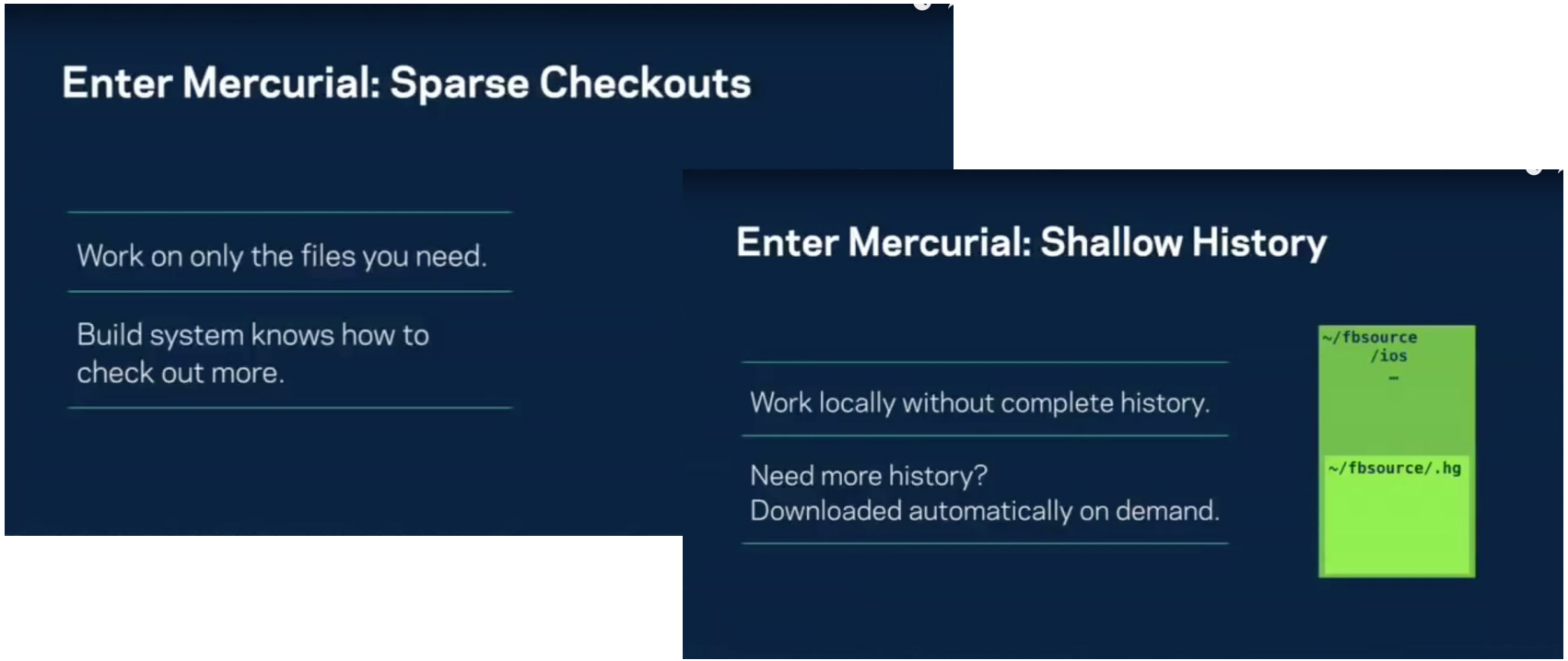
3b. Version control

- Problem: even git can get slow at Facebook scale
  - 1M+ source control commands run per day
  - 100K+ commits per week



3b. Version control

- Solution: redesign version control

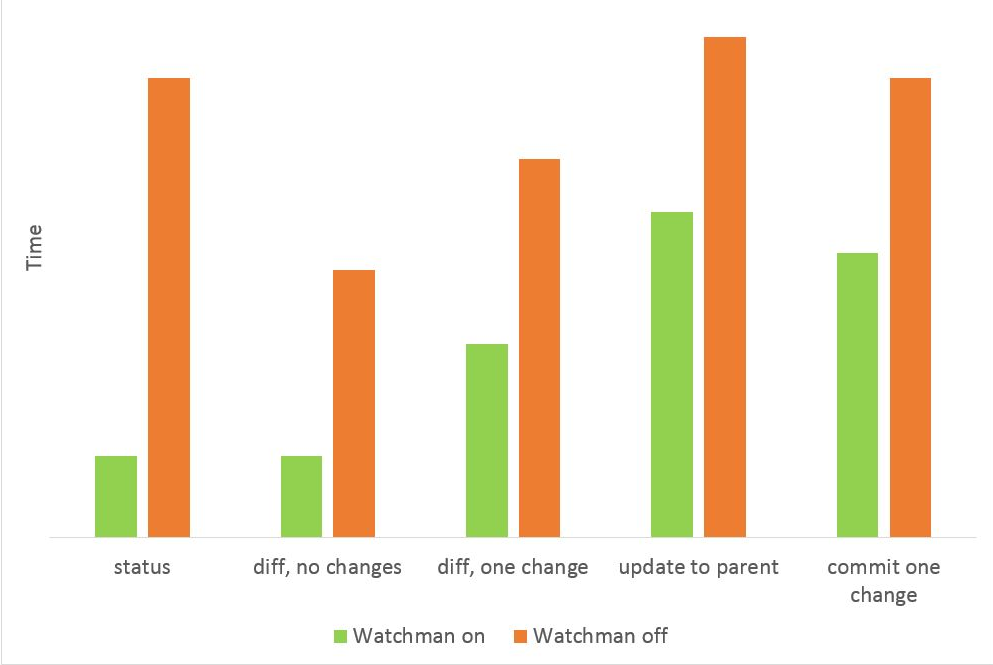


3b. Version control

- Solution: redesign version control
  - Query build system's file monitor, Watchman, to see which files have changed

3b. Version control

- Solution: redesign version control
  - Query build system's file monitor, Watchman, to see which files have changed → **5x faster "status" command**



3b. Version control

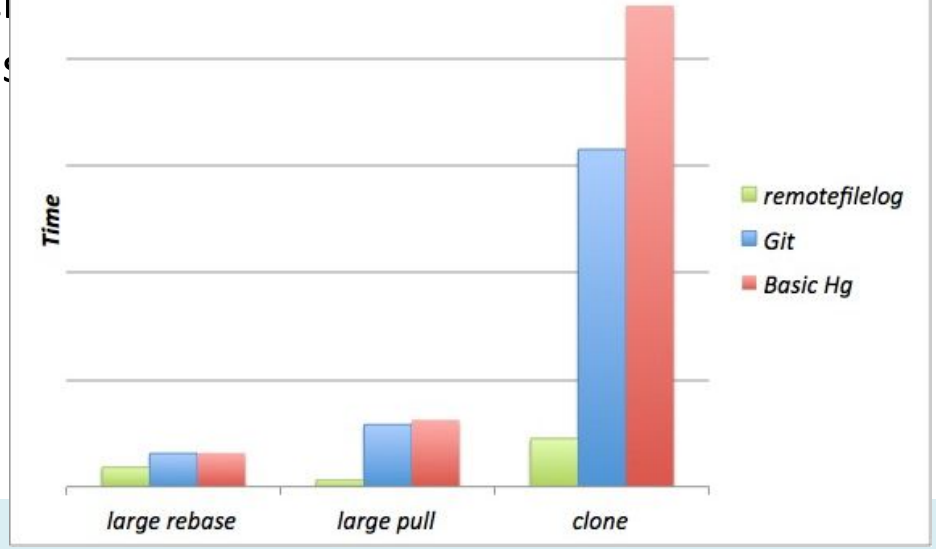
- Solution: redesign version control
  - Sparse checkouts??? (remember, git is a distributed VCS)

3b. Version control

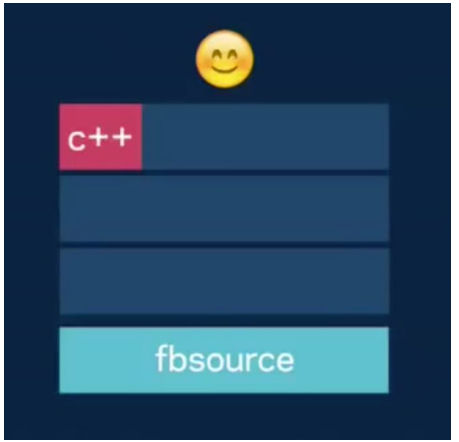
- Solution: redesign version control
  - Sparse checkouts:
  - Change the clone and pull commands to download only the commit metadata, while omitting all file changes (the bulk of the download)
  - When a user performs an operation that needs the contents of files (such as checkout), download the file contents on demand using existing memcache infrastructure

3b. Version control

- Solution: redesign version control
  - Sparse checkouts → **10x faster clones and pulls**
  - Change the clone and pull commands to download only the commit metadata, while omitting all file changes (the bulk of the download)
  - When a user performs an operation that needs the contents of files (such as chec[illegible]nd using existing memcache infra[illegible]



4. Monolithic repository

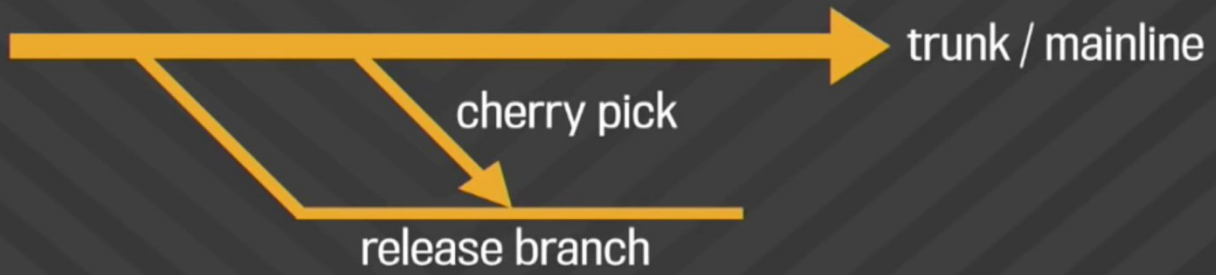


Monolithic repository – no major use of branches for development

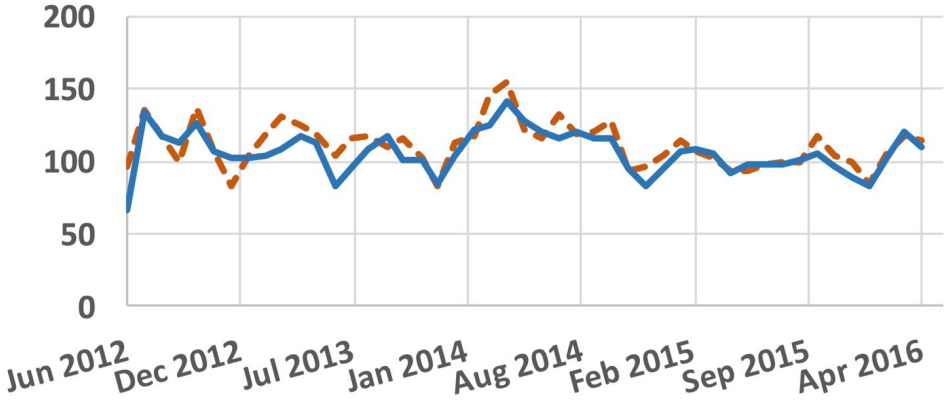
Trunk-based development

Combined with a centralized repository, this defines the monolithic model

- Piper users work at “head”, a consistent view of the codebase
- All changes are made to the repository in a single, serial ordering
- There is no significant use of branching for development
- Release branches are cut from a specific revision of the repository

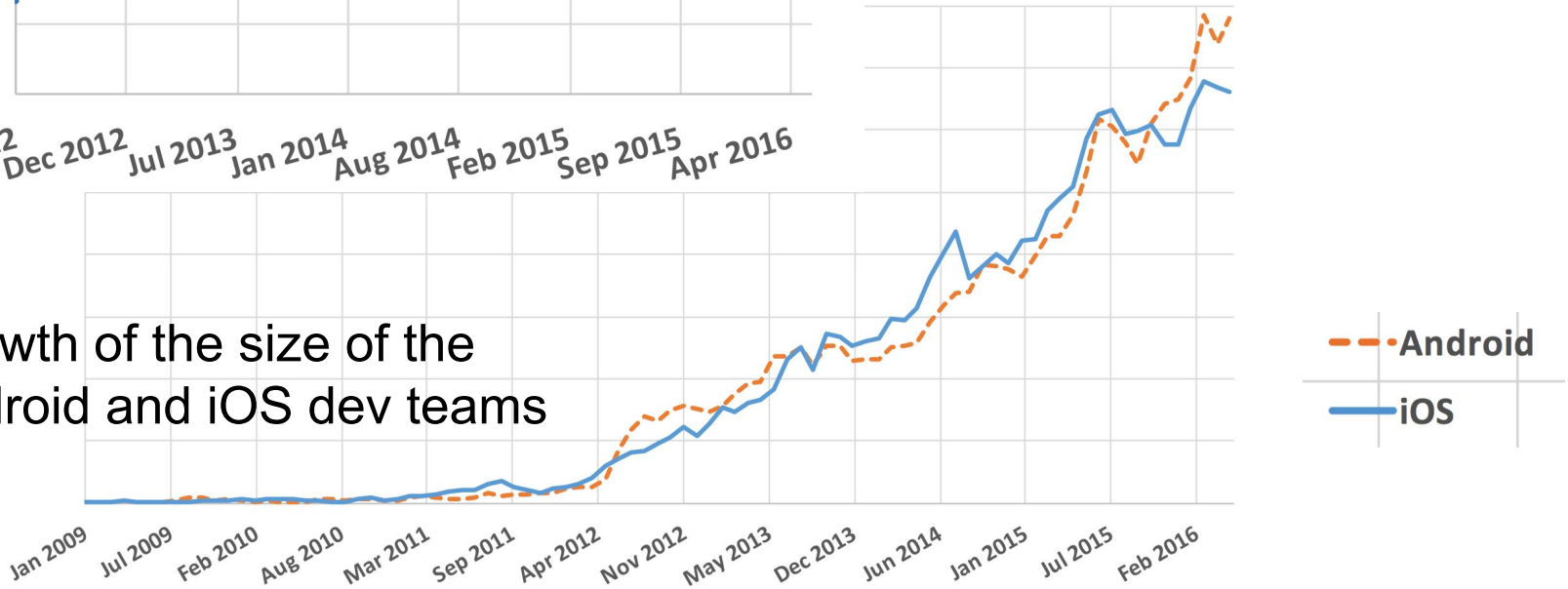


Did it work? Yes. Sustained productivity at Facebook



Lines Committed Per Developer Per Day

Growth of the size of the Android and iOS dev teams



to be continued …

MONOREPO VS MANY REPOS

A recent history of code organization

- A single team with a monolithic application in a single repository

…

- Multiple teams with many separate applications in many separate repositories
- Multiple teams with many ~~separate applications~~ **microservices** in many separate repositories
- A single team with many microservices in many repositories

…

- Many teams with many applications in one big **Monorepo**

What is a Monolithic Repository (monorepo)?

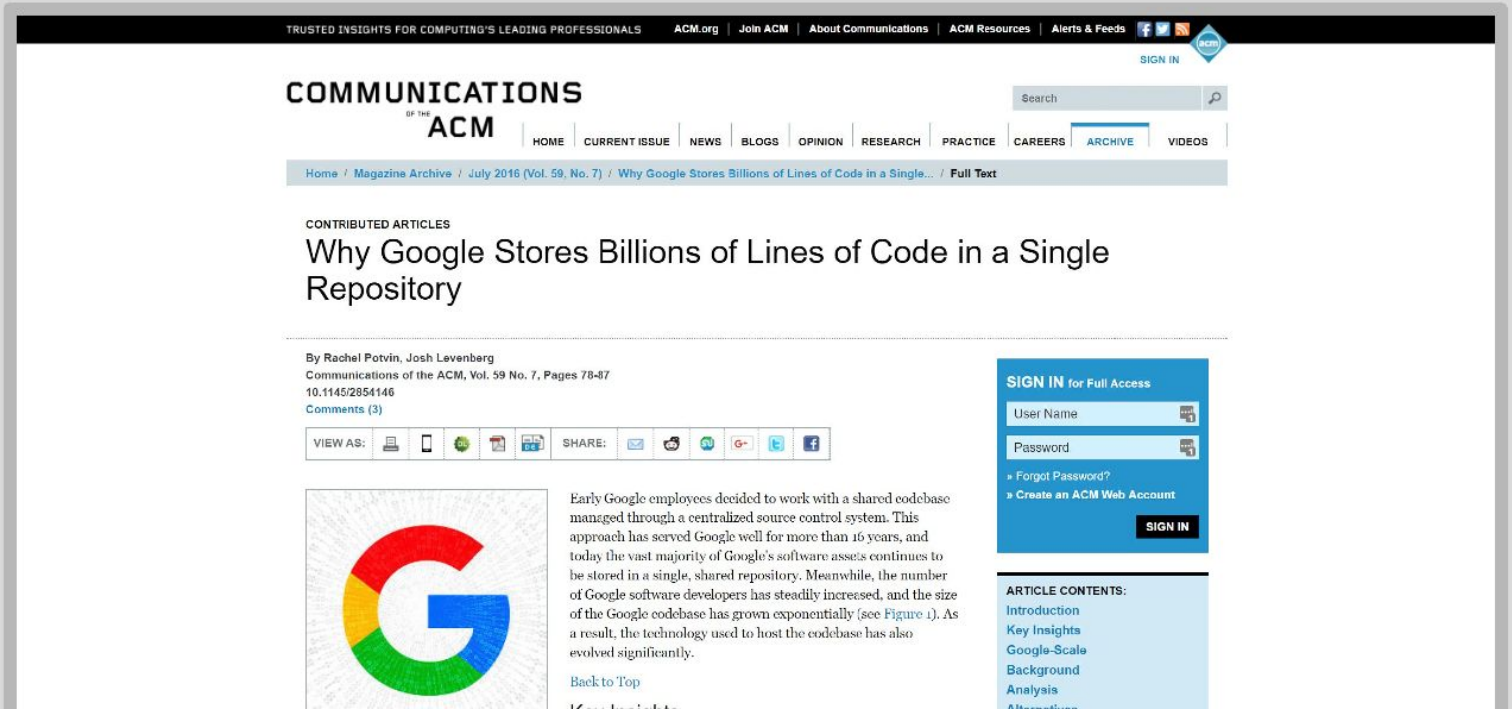
A **single** version control repository containing multiple

- projects
- applications
- libraries,

often using a common build system.

Monorepos in industry

Google (computer science version)



TRUSTED INSIGHTS FOR COMPUTING'S LEADING PROFESSIONALS ACM.org | Join ACM | About Communications | ACM Resources | Alerts & Feeds

SIGN IN

COMMUNICATIONS OF THE ACM

Search

HOME | CURRENT ISSUE | NEWS | BLOGS | OPINION | RESEARCH | PRACTICE | CAREERS | ARCHIVE | VIDEOS

Home / Magazine Archive / July 2016 (Vol. 59, No. 7) / Why Google Stores Billions of Lines of Code in a Single... / Full Text

CONTRIBUTED ARTICLES

Why Google Stores Billions of Lines of Code in a Single Repository

By Rachel Potvin, Josh Levenberg
Communications of the ACM, Vol. 59 No. 7, Pages 78-87
10.1145/2854146
Comments (3)

VIEW AS: SHARE:

Early Google employees decided to work with a shared codebase managed through a centralized source control system. This approach has served Google well for more than 16 years, and today the vast majority of Google's software assets continues to be stored in a single, shared repository. Meanwhile, the number of Google software developers has steadily increased, and the size of the Google codebase has grown exponentially (see Figure 1). As a result, the technology used to host the codebase has also evolved significantly.

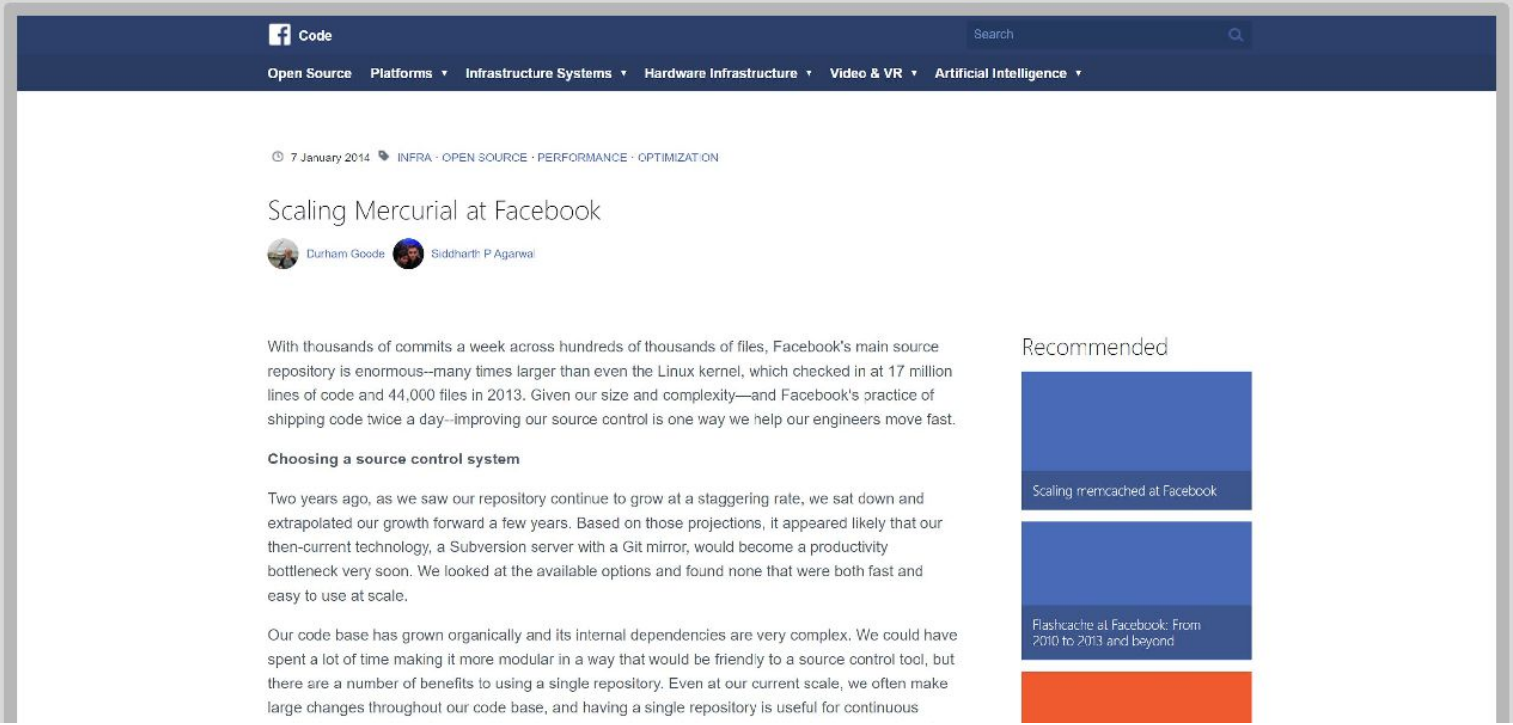
Back to Top

SIGN IN for Full Access
User Name
Password
» Forgot Password?
» Create an ACM Web Account
SIGN IN

ARTICLE CONTENTS:
Introduction
Key Insights
Google-Scale
Background
Analysis

Monorepos in industry

Scaling Mercurial at Facebook



Code

Search

Open Source | Platforms | Infrastructure Systems | Hardware Infrastructure | Video & VR | Artificial Intelligence

7 January 2014 · INFRA · OPEN SOURCE · PERFORMANCE · OPTIMIZATION

Scaling Mercurial at Facebook

Durham Goode · Siddharth P Agarwal

With thousands of commits a week across hundreds of thousands of files, Facebook's main source repository is enormous--many times larger than even the Linux kernel, which checked in at 17 million lines of code and 44,000 files in 2013. Given our size and complexity—and Facebook's practice of shipping code twice a day--improving our source control is one way we help our engineers move fast.

Choosing a source control system

Two years ago, as we saw our repository continue to grow at a staggering rate, we sat down and extrapolated our growth forward a few years. Based on those projections, it appeared likely that our then-current technology, a Subversion server with a Git mirror, would become a productivity bottleneck very soon. We looked at the available options and found none that were both fast and easy to use at scale.

Our code base has grown organically and its internal dependencies are very complex. We could have spent a lot of time making it more modular in a way that would be friendly to a source control tool, but there are a number of benefits to using a single repository. Even at our current scale, we often make large changes throughout our code base, and having a single repository is useful for continuous

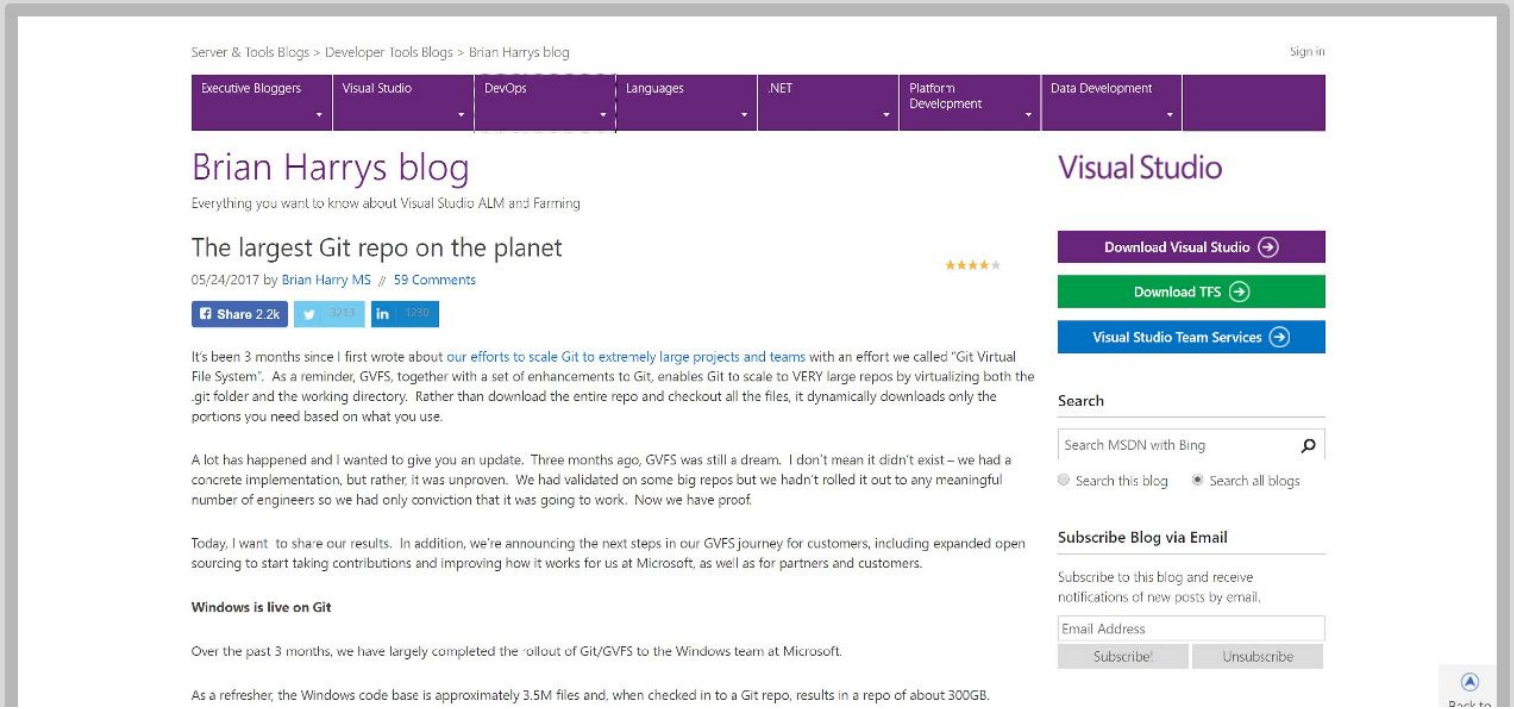
Recommended

Scaling memcached at Facebook

Flashcache at Facebook: From 2010 to 2013 and beyond

Monorepos in industry

Microsoft claim the largest git repo on the planet



Server & Tools Blogs > Developer Tools Blogs > Brian Harrys blog

Sign in

Executive Bloggers | Visual Studio | DevOps | Languages | .NET | Platform Development | Data Development

Brian Harrys blog

Everything you want to know about Visual Studio ALM and Farming

The largest Git repo on the planet

05/24/2017 by Brian Harry MS // 59 Comments

Share 2.2k

It's been 3 months since I first wrote about our efforts to scale Git to extremely large projects and teams with an effort we called "Git Virtual File System". As a reminder, GVFS, together with a set of enhancements to Git, enables Git to scale to VERY large repos by virtualizing both the .git folder and the working directory. Rather than download the entire repo and checkout all the files, it dynamically downloads only the portions you need based on what you use.

A lot has happened and I wanted to give you an update. Three months ago, GVFS was still a dream. I don't mean it didn't exist – we had a concrete implementation, but rather, it was unproven. We had validated on some big repos but we hadn't rolled it out to any meaningful number of engineers so we had only conviction that it was going to work. Now we have proof.

Today, I want to share our results. In addition, we're announcing the next steps in our GVFS journey for customers, including expanded open sourcing to start taking contributions and improving how it works for us at Microsoft, as well as for partners and customers.

Windows is live on Git

Over the past 3 months, we have largely completed the rollout of Git/GVFS to the Windows team at Microsoft.

As a refresher, the Windows code base is approximately 3.5M files and, when checked in to a Git repo, results in a repo of about 300GB.

Visual Studio

Download Visual Studio

Download TFS

Visual Studio Team Services

Search

Search MSDN with Bing

Search this blog | Search all blogs

Subscribe Blog via Email

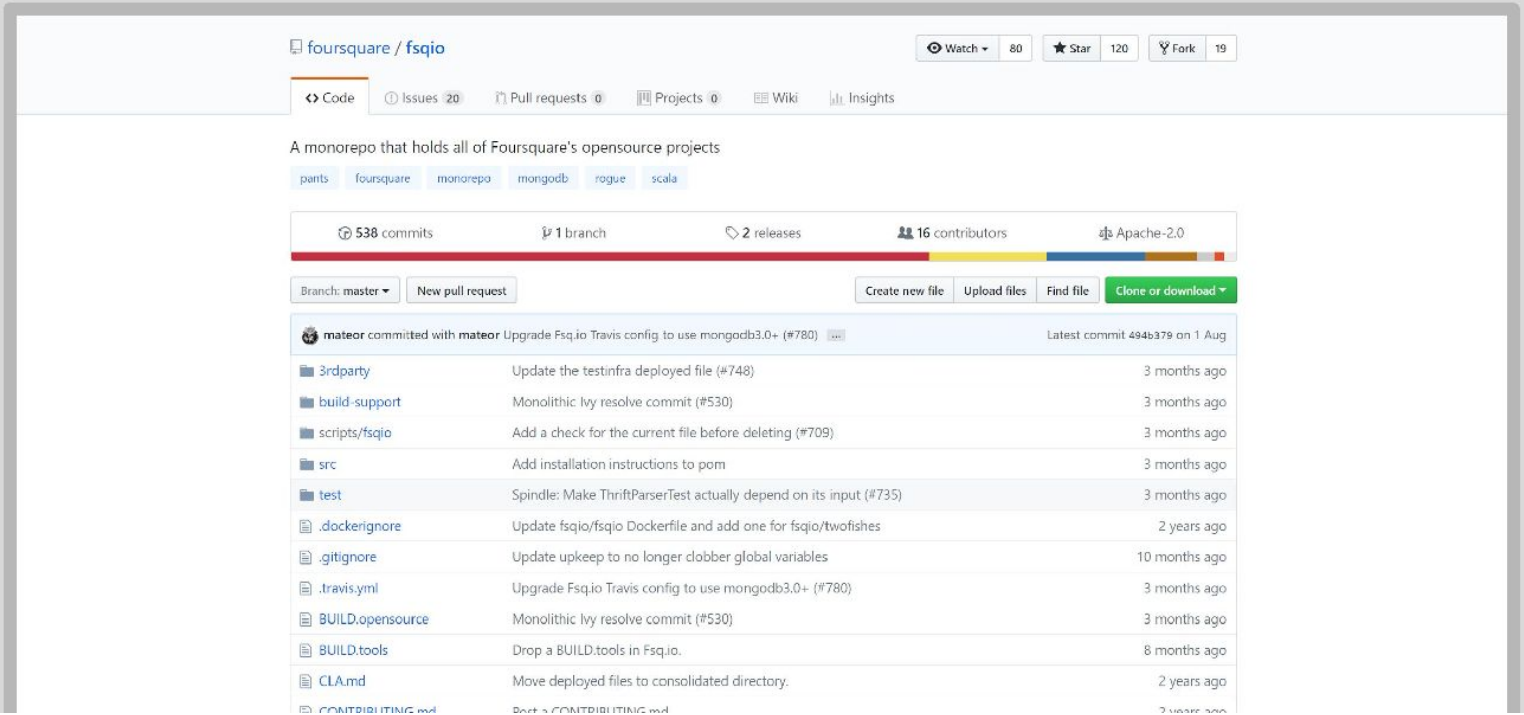
Subscribe to this blog and receive notifications of new posts by email.

Email Address

Subscribe | Unsubscribe

Monorepos in open-source

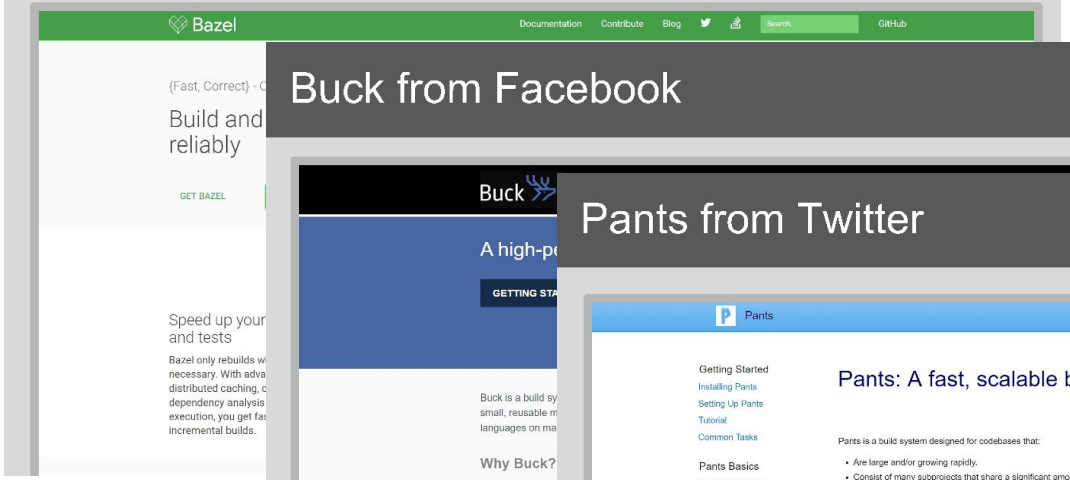
foresquare public monorepo



Monorepos in open-source

Common build system

Bazel from Google



Buck from Facebook

Pants from Twitter

institute for SOFTWARE RESEARCH

Some advantages of monorepos

High Discoverability For Developers

- Developers can read and explore the whole codebase
- `grep`, IDEs and other tools can search the whole codebase
- IDEs can offer auto-completion for the whole codebase
- Code Browsers can links between all artifacts in the codebase

Code-Reuse is cheap

Almost zero cost in introducing a new library

- Extract library code into a new directory/component
- Use library in other components
- Profit!

Refactorings in one commit

Allow large scale refactorings with one single, atomic, history-preserving commit

- Extract Library/Component
- Rename Functions/Methods/Components
- Housekeeping (phpcs-fixer, Namespacing, ...)

Another refactoring example

- Make large backward incompatible changes easily... especially if they span different parts of the project

- For example, old APIs can be removed with confidence
  - Change an API endpoint code **and** all its usages in **all** projects in **one** pull request

Some more advantages

- Easy continuous integration and code review for changes spanning several projects
- (Internal) dependency management is a non-issue
- Less context switching for developers
- Code more reusable in other contexts
- Access control is easy

Some downsides

- Require collective responsibility for team and developers
- Require trunk-based development
  - Feature toggles are technical debt (recall financial services example)
- Force you to have only one version of everything
- Scalability requirements for the repository
- Can be hard to deal with updates around things like security issues
- Build and test bloat without very smart build system
- Slow VCS without very smart system
- Permissions?

Summary

- Version control has many advantages
  - History, traceability, versioning
  - Collaborative and parallel development
- Collaboration with branches
  - Different workflows
- From local to central to distributed version control

Summary

- Configuration management
  - Treat infrastructure as code
  - Git is powerful
- Release management: versioning, branching, …
- Software development at scale requires a lot of infrastructure
  - Version control, build managers, testing, continuous integration, deployment, …
- It’s hard to scale development
  - Move towards heavy automation (DevOps)
- Continuous deployment increasingly common
- Opportunities from quick release, testing in production, quick rollback