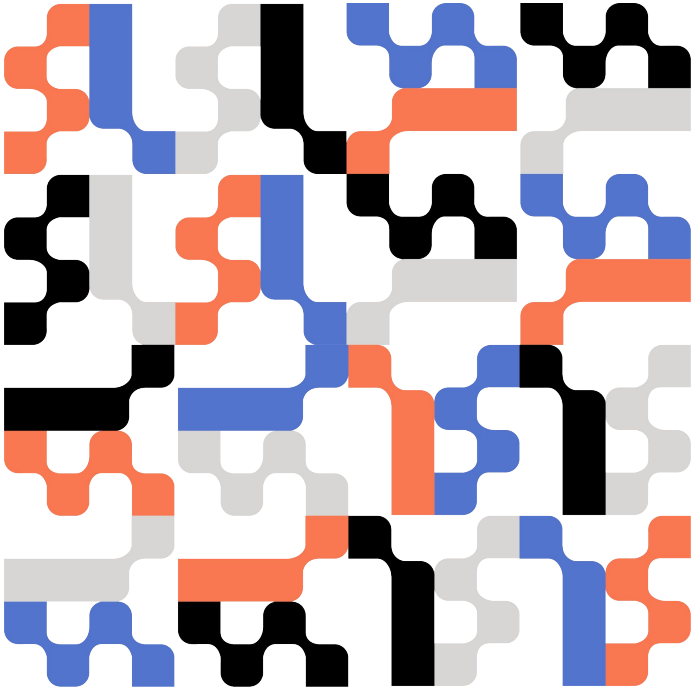
MODUS CREATE | TWEAG by Modus Create

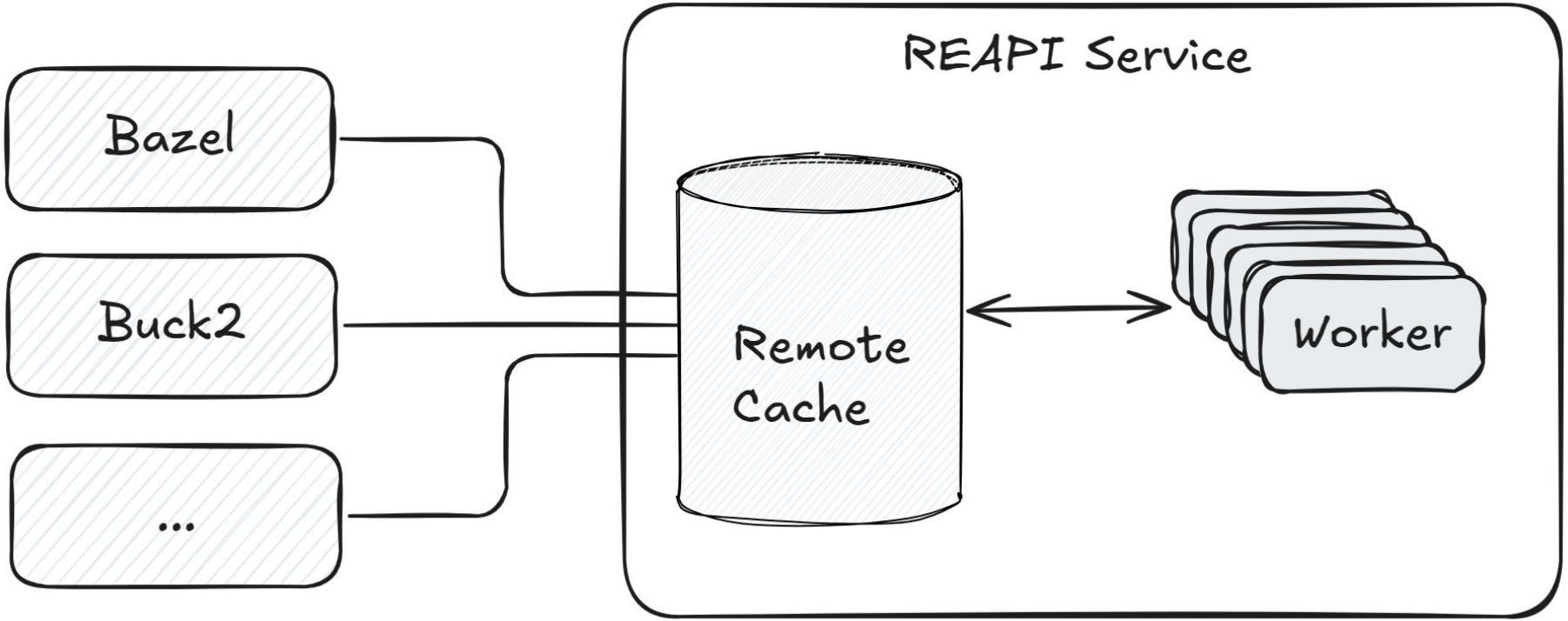
asset-fuse

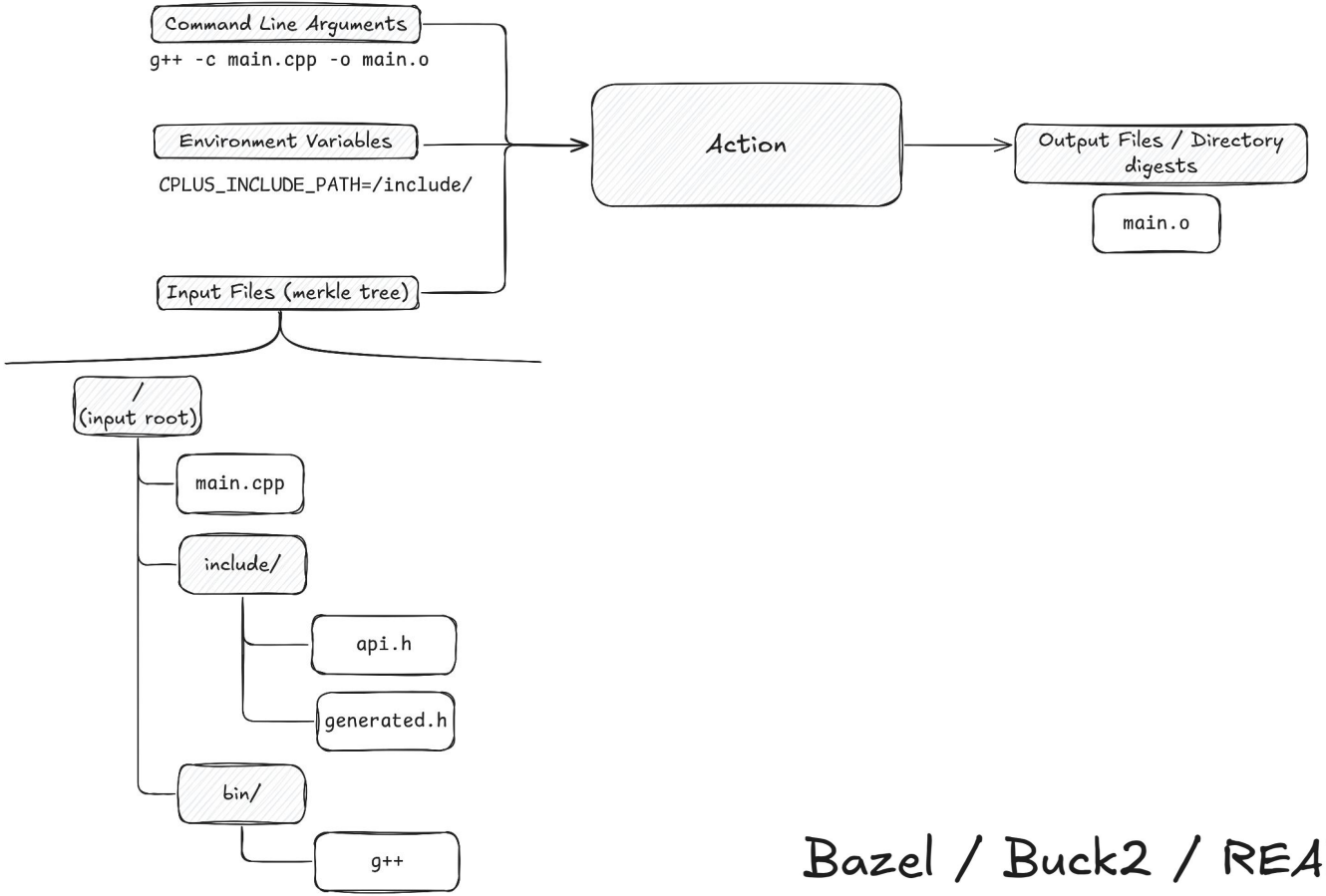
Bringing Large Files to Buck2 and Bazel



07.Apr.2025

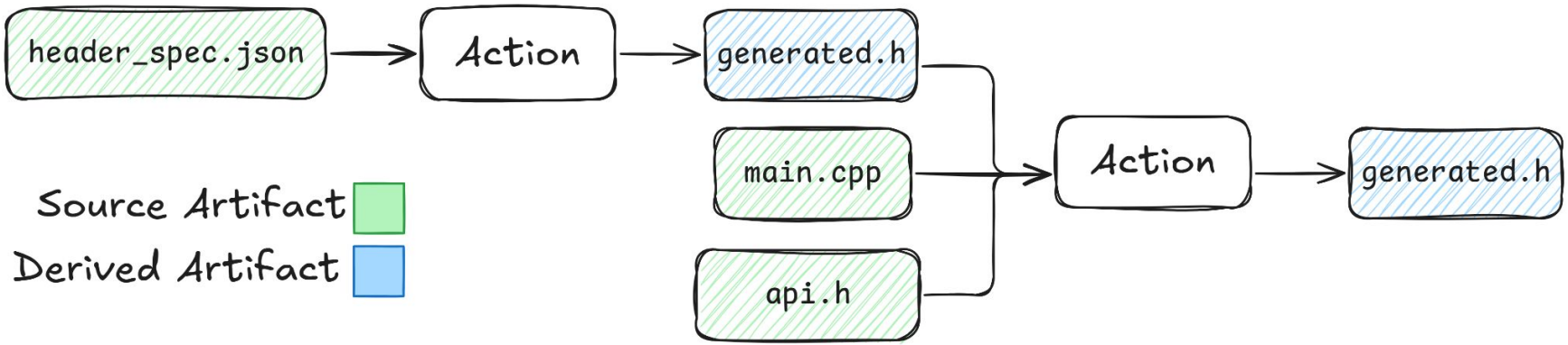
Typical Setup



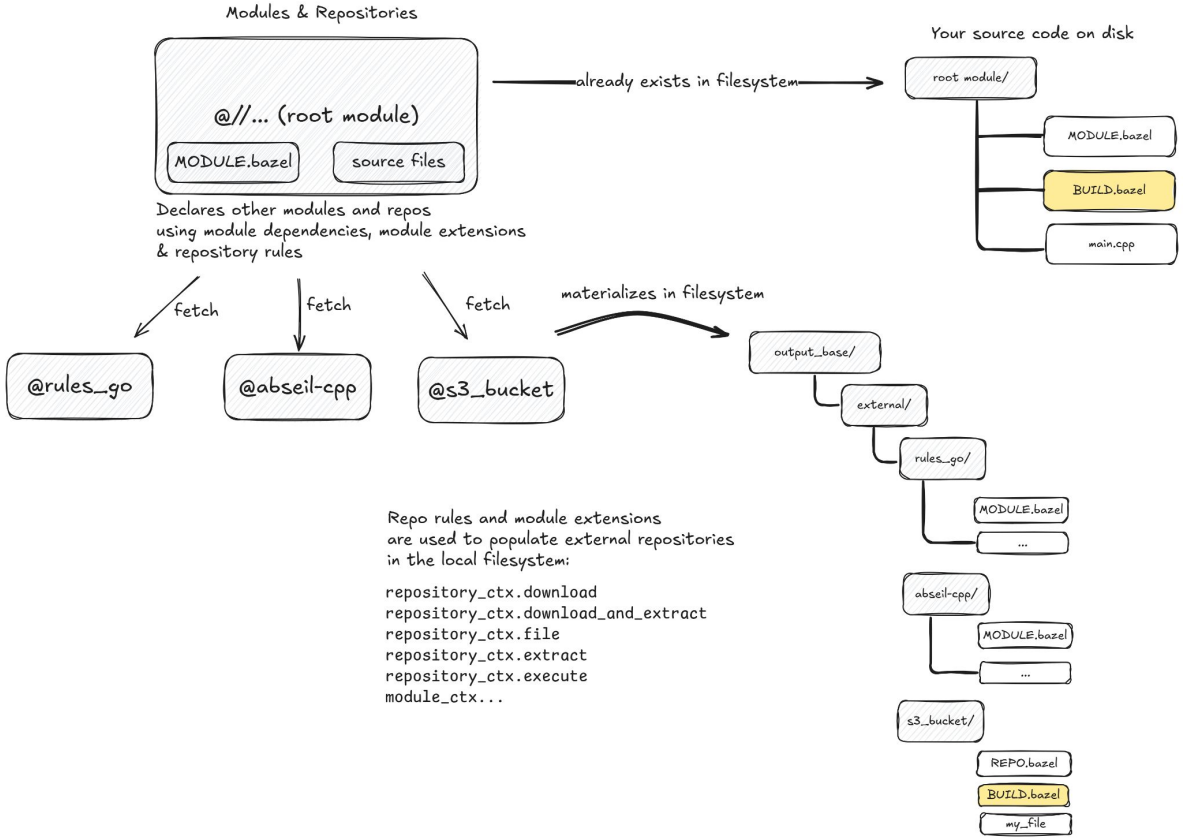


Bazel / Buck2 / REAPI Actions

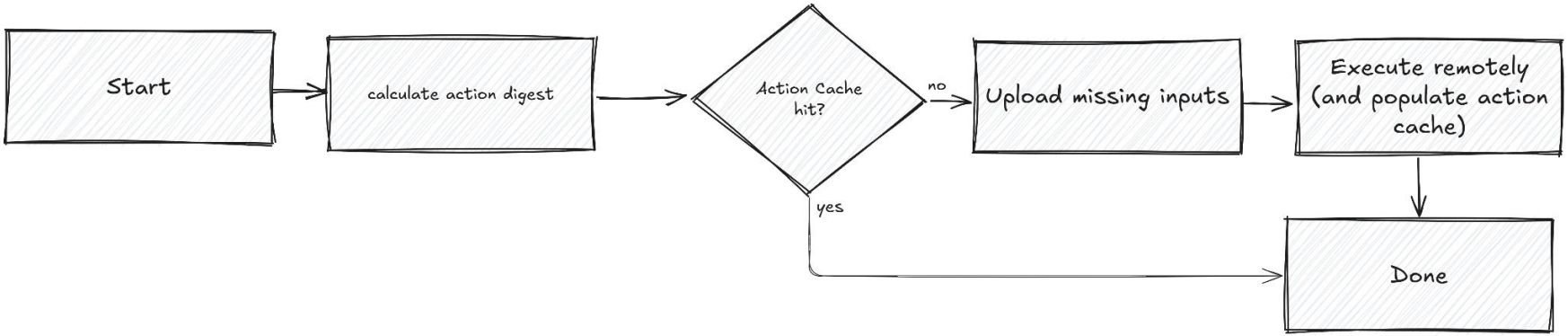
Path from sources to outputs



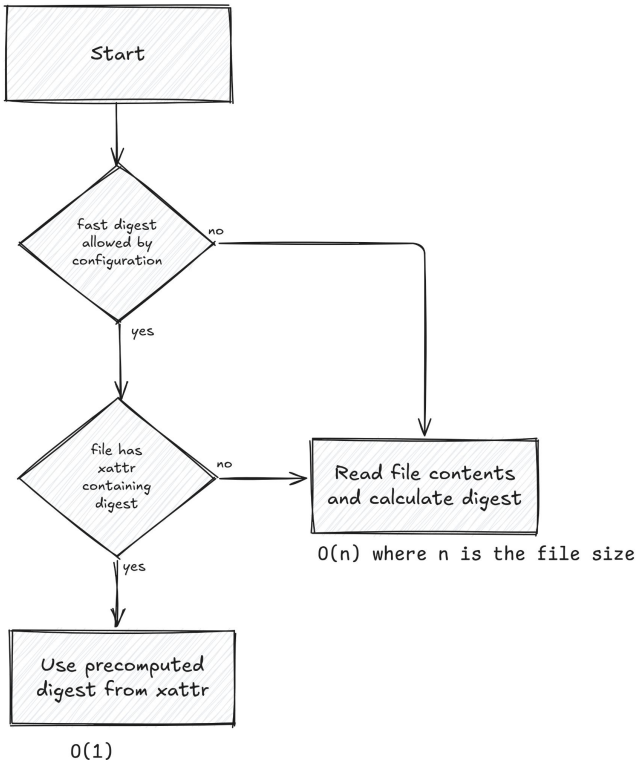
Bazel's Loading Phase



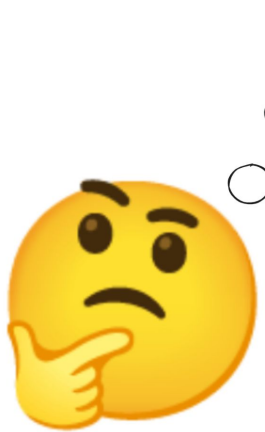
Running an action remotely



Computing Digests of Source Artifacts



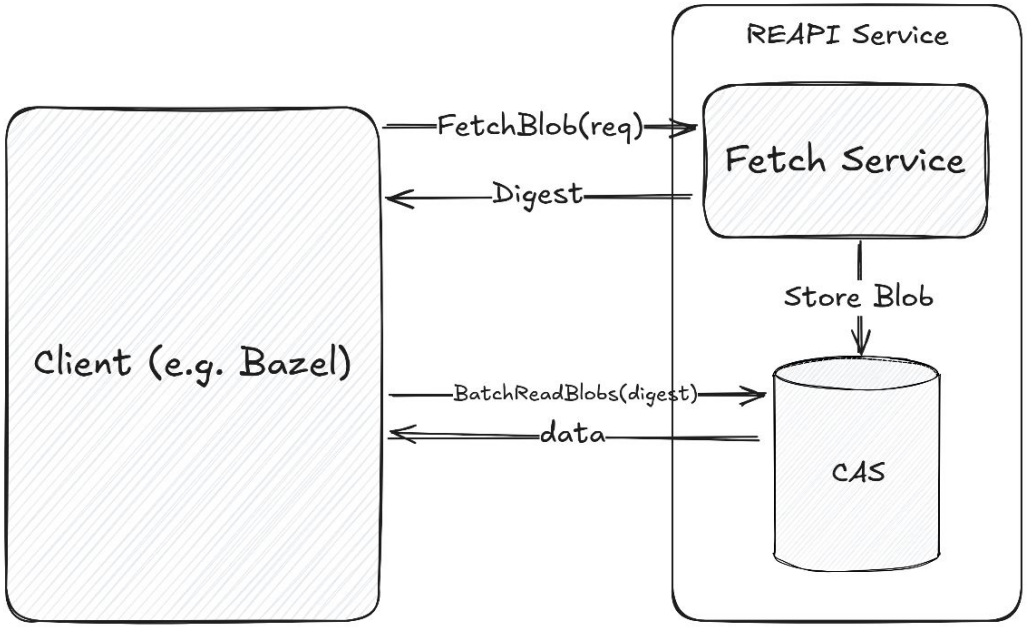
The file doesn't even need to exist on disk, as long as the filesystem returns the correct size via stat and digest via getxattr.

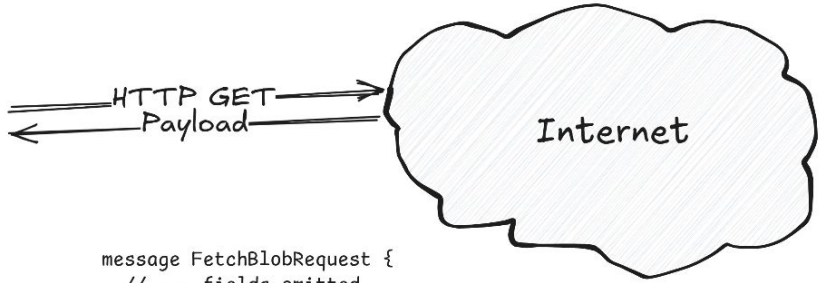


Problem Statement

- Bazel eagerly materializes whole repositories on disk (repo rules)
- “Build-without-the-Bytes” doesn’t work for downloaded files
- BUILD files needed locally for analysis
- Source files (data) only needed remotely

Remote-Asset API

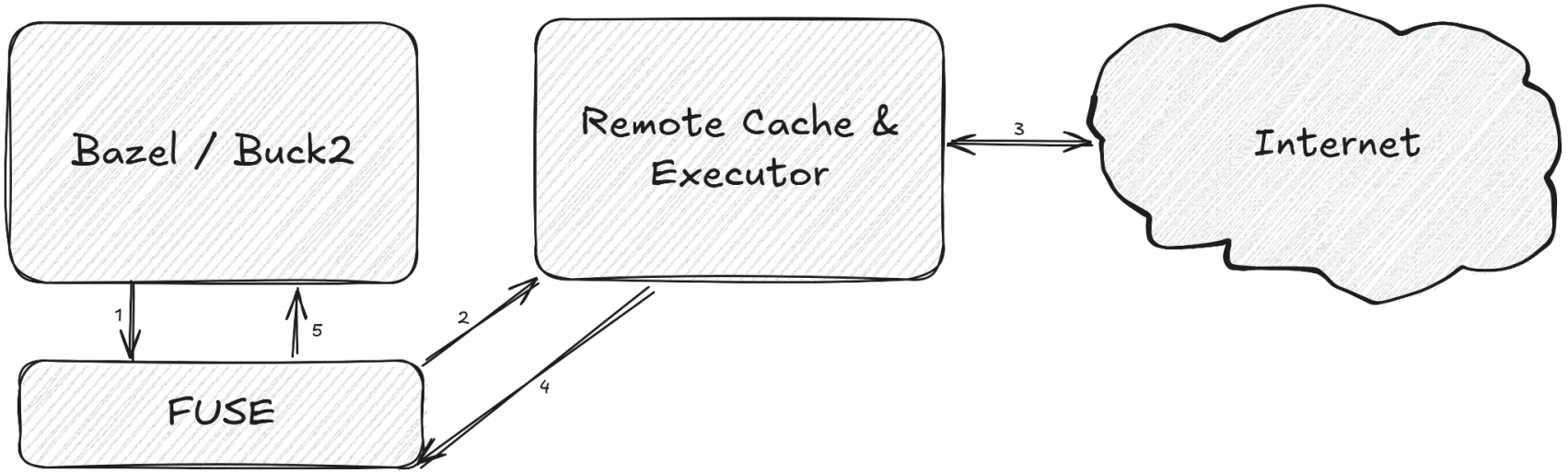





```
message FetchBlobRequest {
  // ... fields omitted
  repeated string uris = 4;
  repeated Qualifier qualifiers = 5;
  // ...
}

message FetchBobResponse {
  // ... fields omitted
  build.bazel.remote.execution.v2.Digest blob_digest = 5;
  // ...
}
```

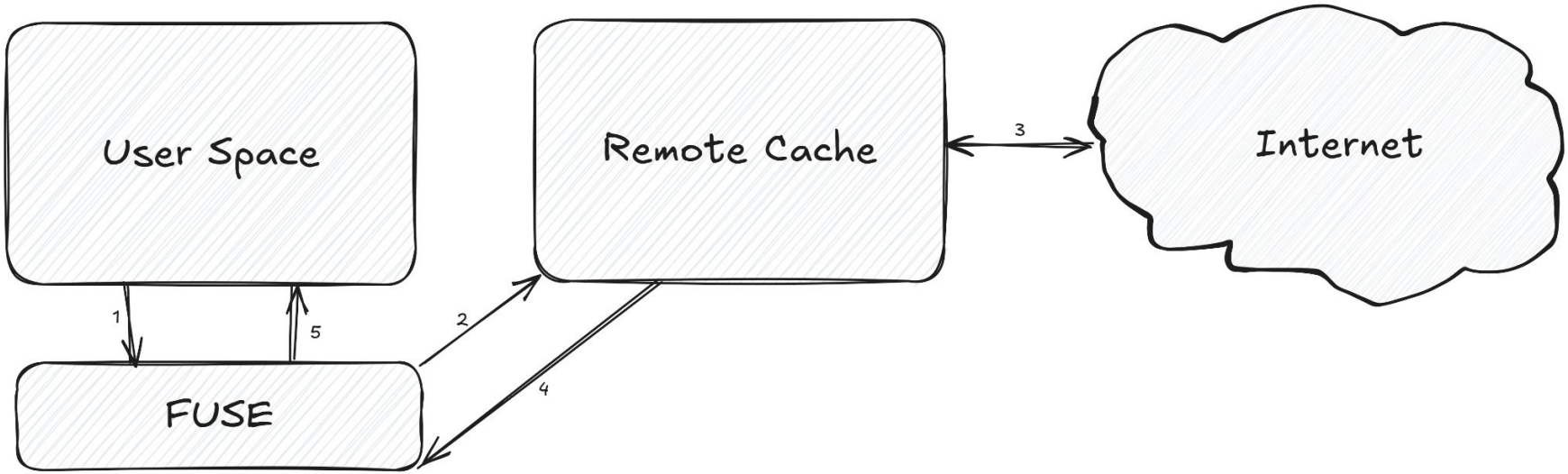
Architecture



Manifest of contents

```
{
  "paths": {
    "content/logo.jpg": {
      "integrity": "sha256-NmL...oPMGeXvYrcUNs=",
      "size": 72281
    },
    "content/background.png": {
      "integrity": "sha256-4Zh...sJRSJCGMMr8v48=",
      "size": 19180
    }
  },
  "uri_templates": [
    "https://mir-s3-cdn-cf.behance.net/project_modules/max_1200/{basename}",
    "https://ghcr.io/v2/malt3/asset-fuse-cas/blobs/sha256:{sha256}"
  ]
}
```

Reading a file



“git-lfs on steroids”

git-lfs stores pointer files in Git

```
version https://git-lfs.github.com/spec/v1
oid sha256:f0d4...6532f4ba49b
size 12772
```

DVC stores .dvc files in Git

```
md5: f5ea021eddd7b1df6de80b904cba1da6
frozen: true
deps:
- path: get-started/data.xml
  repo:
    url: https://github.com/iterative/dataset-registry
    rev_lock: f59388cd04276e75d70b2136597aaa27e7937cc3
outs:
- md5: 22a1a2931c8370d3aeedd7183606fd7f
  size: 14445097
  hash: md5
  path: data.xml
```

“git-lfs on steroids”

asset-fuse stores a manifest in Git

```
{
  "paths": {
    "content/logo.jpg": {
      "integrity": "sha256-NmL...oPMGeXvYrcUNs=",
      "size": 72281
    },
    "content/background.png": {
      "integrity": "sha256-4Zh...sJRSJCGMMr8v48=",
      "size": 19180
    }
  }
  "uri_templates": [
    "https://mir-s3-cdn-cf.behance.net/project_modules/max_1200/{basename}",
    "https://ghcr.io/v2/malt3/asset-fuse-cas/blobs/sha256:{sha256}"
  ]
}
```

“git-lfs on steroids”

git-lfs downloads files eagerly

```
$ git pull
Updating
91f2a01b..3345073c
...
```

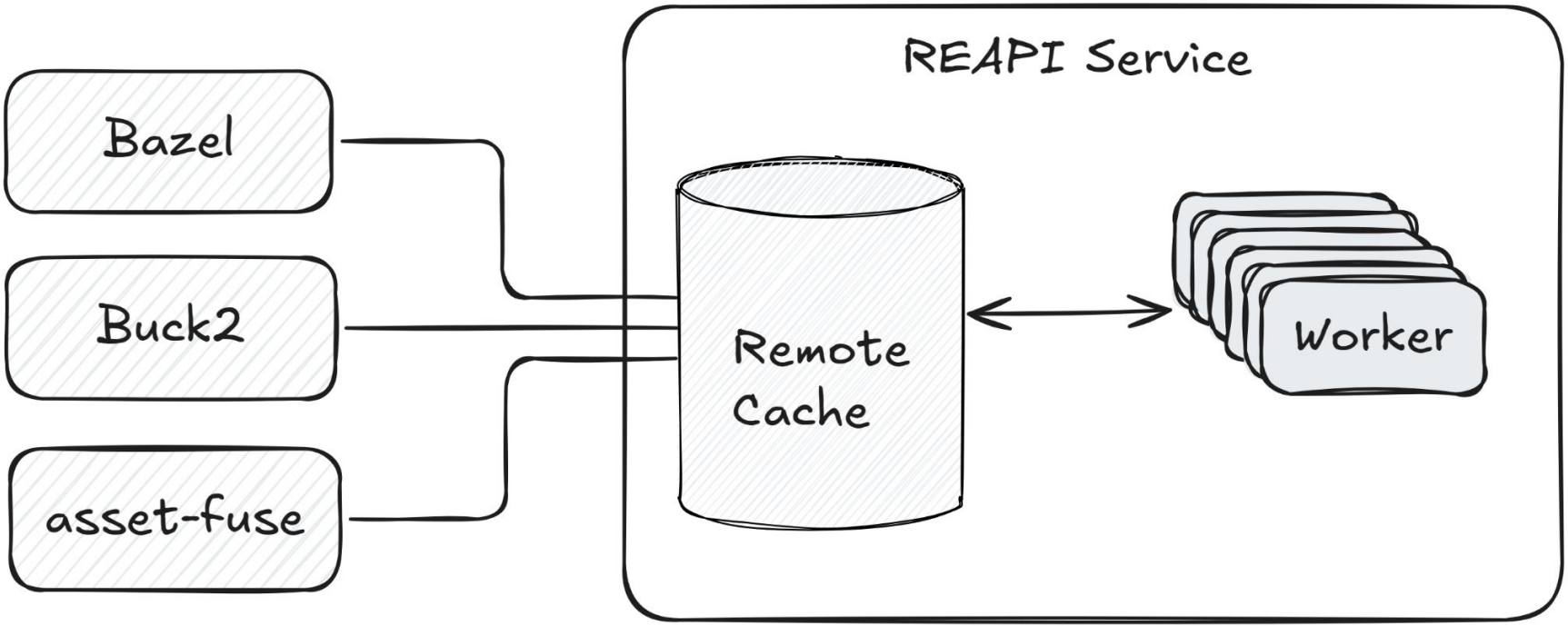
DVC downloads files eagerly

```
$ dvc pull
Downloading ...
```

asset-fuse is lazy

```
$ asset-fuse mount mnt
$ open mnt/content/logo.jpg
```

asset-fuse + Bazel/Buck2 + RBE



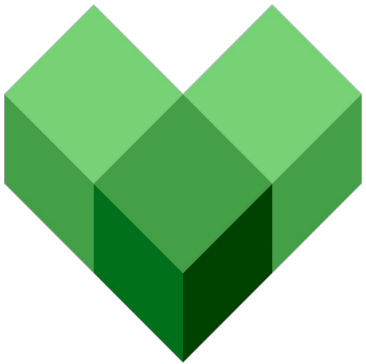
Benefits

- Optimal: Large files can be prefetched into remote cache
- Minimal: Large files don’t need to be materialized on your laptop

⇒ You can work with enormous files with ease

Standing on the shoulders of giants

- Google: SrcFs is a custom, internal FUSE filesystem that exposes digests via xattr for Blaze / Bazel
- Meta: EdenFs is a custom, open source FUSE filesystem that exposes digests via xattr for Buck2
- BuildBarn uses FUSE extensively to optimize file access for remote execution and via bb-clientd



What's next?

- `FetchDirectory`
  - Remote extraction of tar files
  - Mount Git repository
- Better local caching
- More data sources
  - Mount git-lfs data
  - Parse DVC files
  - Format for large manifests (sqlite?)

Reach out!

- Working beta
- Looking for early adopters and feedback

github.com/tweag/asset-fuse

TWEAG
by Modus Create

THANK YOU!

MODUS CREATE | TWEAG by Modus Create