Caching and consistency: Network Files Systems + the Web

COS 418/518: *Distributed Systems*
Lecture 3

Michael Freedman

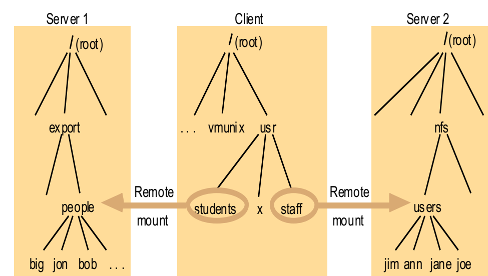
Abstraction, abstraction, abstraction!

- Local file systems
  - Disks are terrible abstractions: low-level blocks, etc.
  - Directories, files, links much better
- Distributed file systems
  - Make a remote file system look local
  - Today: NFS (Network File System)
    - Developed by Sun in 1980s, still used today!
- Web servers
  - Make remote content look local

3 Goals: Make operations appear

- Local
- Consistent
- Fast

NFS: Naming indirection, abstraction



"Mount" remote FS (host:path) as local directories

Local FS / Virtual File System (VFS)

`fd = open("path", flags)`

`read(fd, buf, n)`

`write(fd, buf, n)`

`close(fd)`

Computer maintains state that maps `fd` to inode, offset

5

Stateless NFS: Strawman 1

~~fd = open("path", flags)~~

`read("path", buf, n)`

`write("path", buf, n)`

~~close(fd)~~

6

Stateless NFS: Strawman 2

~~fd = open("path", flags)~~

`read("path", offset, buf, n)`

`write("path", offset, buf, n)`

~~close(fd)~~

7

Embed pathnames in syscalls?

	Program 1 on client 1	Program 2 on client 2
1	CHDIR ("dir1")	
2	*fd* ← OPEN ("f", READONLY)	
3		RENAME ("dir1", "dir2")
4		RENAME ("dir3", "dir1")
5	READ (*fd*, *buf*, *n*)	

Time ↓

- Should read refer to current `dir1/f` or `dir2/f` ?
- In UNIX, it's `dir2/f`. How do we preserve in NFS?

8

Stateless NFS (for real)

```
fh = lookup("path", flags)

read(fh, offset, buf, n)

write(fh, offset, buf, n)

getattr(fh)
```

Implemented as Remote Procedure Calls (RPCs)

NFS File Handles (`fh`)

- Opaque identifier provider to client from server
- Includes all info needed to identify file/object on server

volume ID | inode # | generation #

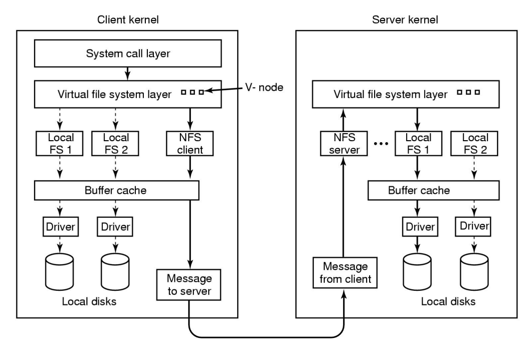
- It's a trick: "store" server state at the client!

NFS File Handles (and versioning)

	Program 1 on client 1	Program 2 on client 2	Time
1		*fd* ← OPEN ("f", READONLY)	↓
2	UNLINK ("f")		
3	*fd* ← OPEN ("f", CREATE)		
4		READ (*fd*, *buf*, *n*)	

- With generation #'s, client 2 continues to interact with "correct" file, even while client 1 has changed "f"
- This versioning appears in many contexts, e.g., MVCC (multiversion concurrency control) in DBs

Are remote == local?



TANSTANFL
(There ain't no such thing as a free lunch)

- With local FS, `read` sees data from "most recent" `write`, even if performed by different process
  - "Read/write coherence", linearizability
- Achieve the same with NFS?
  - Perform all reads & writes synchronously to server
  - Huge cost: high latency, low scalability
- And what if the server doesn't return?
  - Options: hang indefinitely, return ERROR

13

Caching GOOD

Lower latency, better scalability

Consistency HARDER

No longer one single copy of data, to which all operations are serialized

14

Caching options

- Read-ahead: Pre-fetch blocks before needed
- Write-through: All writes sent to server
- Write-behind: Writes locally buffered, send as batch
- Consistency challenges:
  - When client writes, how do others caching data get updated? (Callbacks, …)
  - Two clients concurrently write? (Locking, overwrite, …)

NFS

- Stateless protocol
  - Recovery easy: crashed == slow server
  - Messages over UDP (unencrypted)
- Read from server, caching in NFS client
- NFSv2 was write-through (i.e., synchronous)
- NFSv3 added write-behind
  - Delay writes until `close` or `fsync` from application

16

Exploring the consistency tradeoffs

- Write-to-read semantics too expensive
  - Give up caching, require server-side state, or …
- Close-to-open "session" semantics
  - Ensure an ordering, but only between application `close` and `open`, not all `writes` and `reads`.
  - If B opens after A closes, will see A's writes
  - But if two clients open at same time? No guarantees
    - And what gets written? "Last writer wins"

17

NFS Cache Consistency

- Recall challenge: Potential concurrent writers
- Cache validation:
  - Get file's last modification time from server: `getattr(fh)`
  - Both when first open file, then poll every 3-60 seconds
    - If server's last modification time has changed, flush dirty blocks and invalidate cache
- When reading a block
  - Validate: (current time – last validation time < threshold)
  - If valid, serve from cache. Otherwise, refresh from server

18

Some problems…

- "Mixed reads" across version
  - A reads block 1-10 from file, B replaces blocks 1-20, A then keeps reading blocks 11-20.
- Assumes synchronized clocks. Not really correct.
  - We'll learn about the notion of logical clocks later
- Writes specified by offset
  - Concurrent writes can change offset
  - More on this later with techniques for conflict resolution

19

When statefulness helps

Callbacks
Locks + Leases

20

Locks

- A client can request a lock over a file / byte range
  - Advisory: Well-behaved clients comply
  - Mandatory: Server-enforced
- Client performs writes, then unlocks
- Problem: What if the client crashes?
  - Solution: Keep-alive timer: Recover lock on timeout
    - Problem: what if client alive but network route failed?
      Client thinks has lock, server gives lock to other: "Split brain"

22

Leases

- Client obtains ***lease*** on file for read or write
  - "A lease is a ticket permitting an activity; the lease is valid until some expiration time."
- Read lease allows client to cache clean data
  - *Guarantee:* no other client is modifying file
- Write lease allows safe delayed writes
  - Client can locally modify than batch writes to server
  - *Guarantee:* no other client has file cached

Using leases

- Client requests a lease
  - May be implicit, distinct from file locking
  - Issued lease has file version number for cache coherence
- Server determines if lease can be granted
  - *Read leases* may be granted concurrently
  - *Write leases* are granted exclusively
- If conflict exists, server may send *eviction* notices
  - Evicted write lease must write back
  - Evicted read leases must flush/disable caching
  - Client acknowledges when completed

24

Bounded lease term simplifies recovery

- Before lease expires, client must *renew* lease
- Client fails while holding a lease?
  - Server waits until the lease expires, then unilaterally reclaims
  - If client fails during eviction, server waits then reclaims
- Server fails while leases outstanding? On recovery:
  - Wait *lease period + clock skew* before issuing new leases
  - Absorb renewal requests and/or writes for evicted leases

Statelessness: Web caching

HTTP Caching

- Clients (and proxies) cache documents
  - When should origin be checked for changes?
  - Every time? Every session? Date?
- HTTP includes caching information in headers
  - HTTP 0.9/1.0 used: “Expires: <date>”; “Pragma: no-cache”
  - HTTP/1.1 has “Cache-Control”
    - “No-Cache”, “Max-age: <seconds>”
    - “E-tag: <opaque value>

HTTP Caching

- If not expired: use cached copy
- If expired, use condition GET request to origin
  - “If-Modified-Since: <date>”, “If-None-Match: <etag>”
  - 304 (“Not Modified”) or 200 (“OK”) response

```
GET / HTTP/1.1
Host: sns.cs.princeton.edu
Connection: Keep-Alive
If-Modified-Since: Tue, 1 Feb 2011 …
If-None-Match: "7a11f-10ed-3a75ae4a"
```

```
HTTP/1.1 304 Not Modified
Date: Wed, 02 Feb 2011 ….
Server: Apache/2.2.3 (CentOS)
ETag: "7a11f-10ed-3a75ae4a"
Accept-Ranges: bytes
```

Statefulness: Summary

Server maintain per-client state?
(which files open for reading/writing, what cached, …)

Stateful

- Pros
  - Smaller requests
  - Simpler req processing
  - Better cache coherence, file locking, etc.
- Cons
  - Per-client state limits scalability
  - Fault-tolerance on state required for correctness

Stateless

- Pros
  - Easy server crash recovery
  - No open/close needed
  - Better scalability
- Cons
  - Each request must be fully self-describing
  - Consistency is harder, e.g., no simple file locking

It's all about the state, 'bout the state, …

- Hard state: Don't lose data
  - Durability: State not lost
    - Write to disk, or cold remote backup
    - Exact replica or recoverable (DB: checkpoint + op log)
  - Availability (liveness): Maintain online replicas
- Soft state: Performance optimization
  - Traditionally: Lose at will
  - More recently: Yes for correctness (safety), but how does recovery impact availability (liveness)?

31