redo logging (finish) / distributed systems 1

last time (1)

block groups — keep related data+metadata in one part of disk

- preference, not requirement — exceptions can span multiple block groups
- divide up block/inode indices between block groups

small files: fragments — dividing blocks into pieces

large files: extents — ranges instead of single block pointers

cost of fragments and extents

- complicate block allocation, free block tracking

last time (2)

redo logging

goal: perform multiple updates "at once" (consistency!)

record intention in log

record committing to that intention

at this point: operation "done" for application's perspective
(i.e. OS won't forget about the operation even if crash)

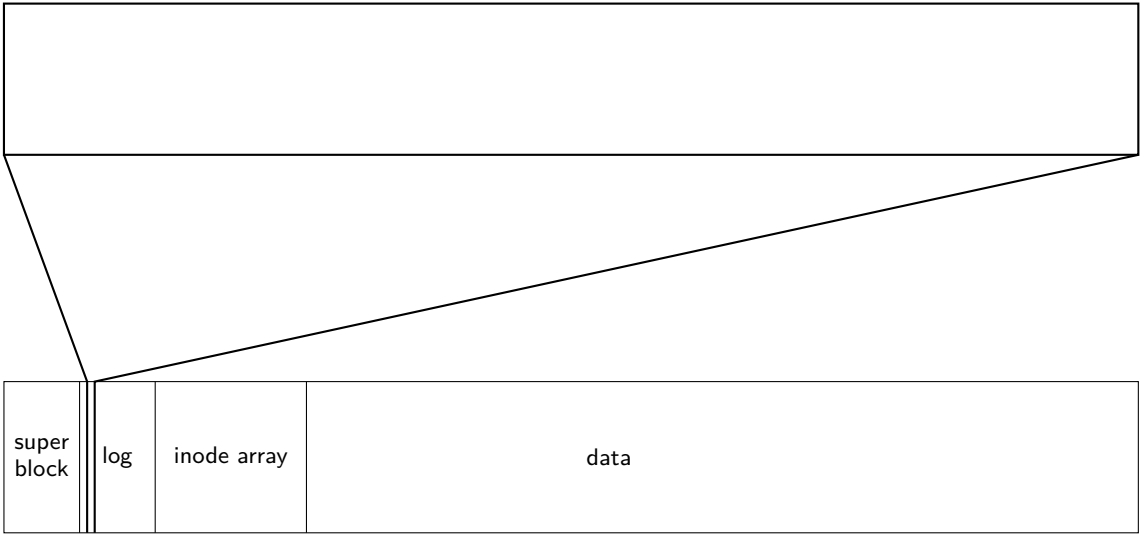
actually do what was intended

on crash: redo what was intended

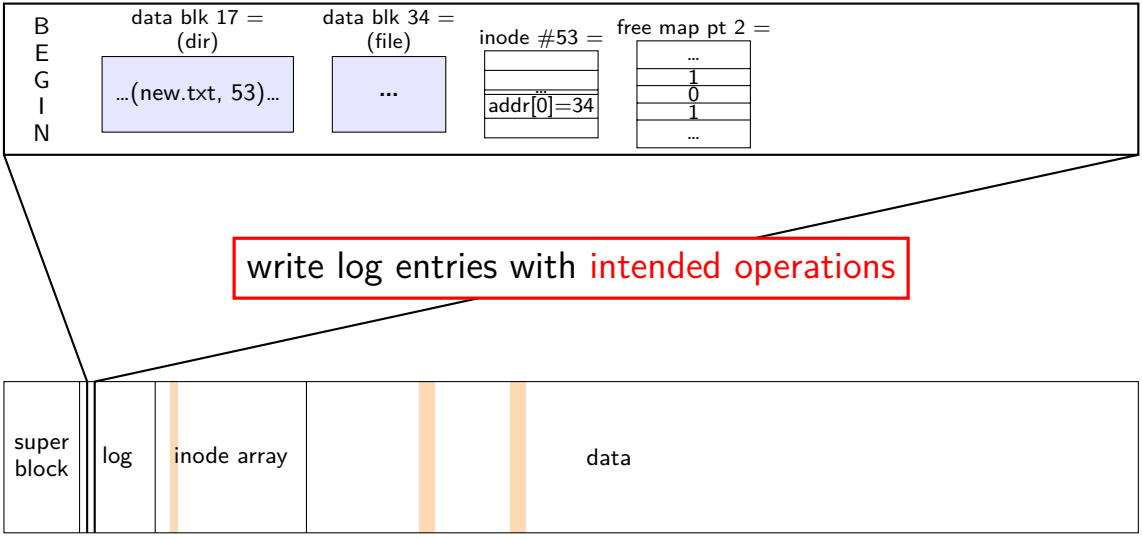
may or may not be repeating operations

eventually: clear log of fully complete operations

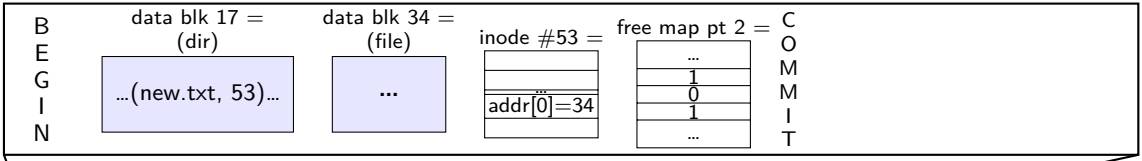
redo logging: file creation



redo logging: file creation

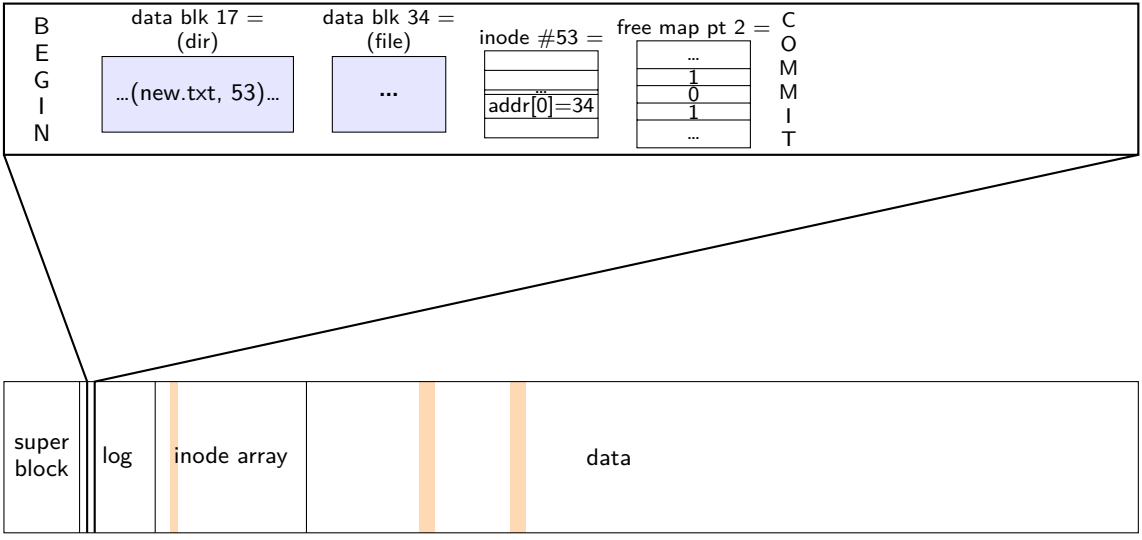


redo logging: file creation

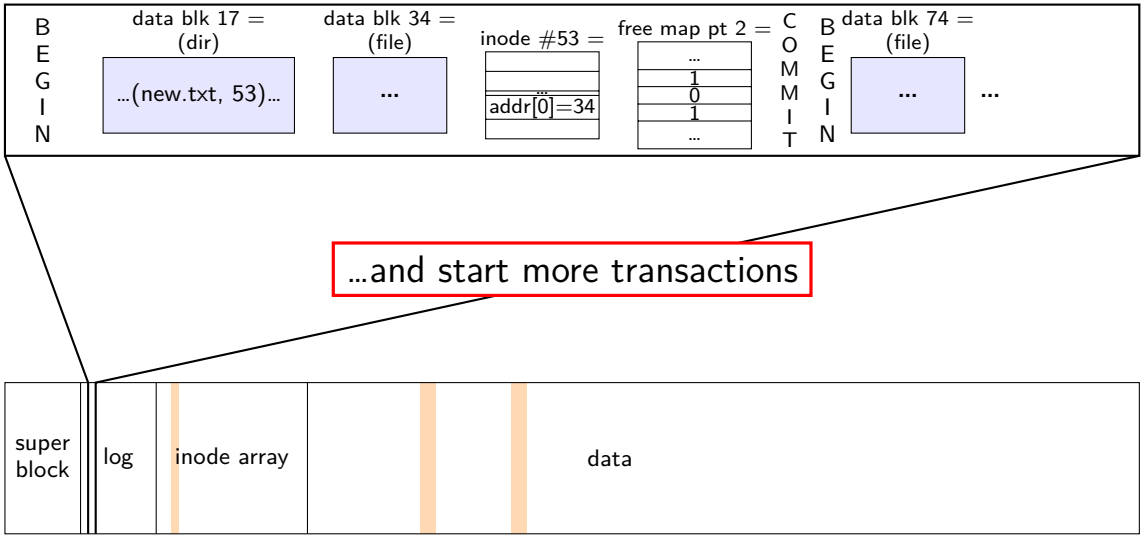


filesystem needs to ensure that committed updates will definitely happen!
mechanism: check this log for commit messages later, and redo them (just in case)

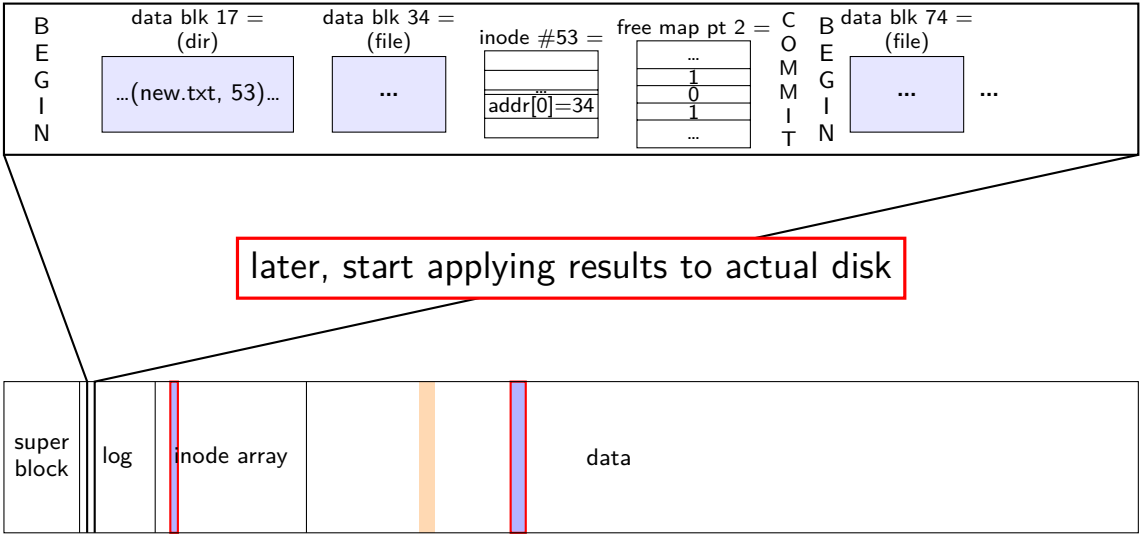
redo logging: file creation



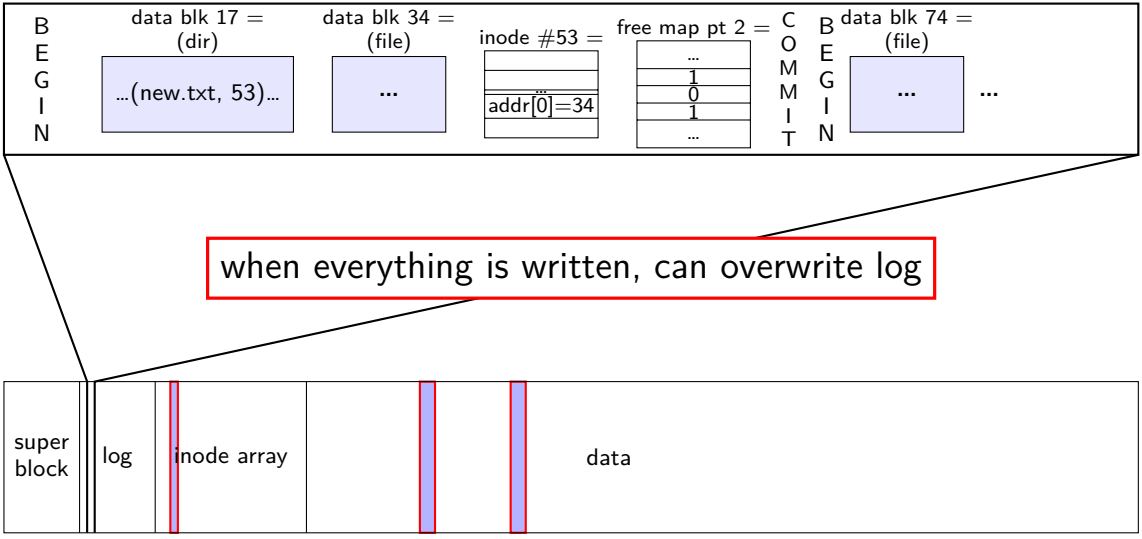
redo logging: file creation



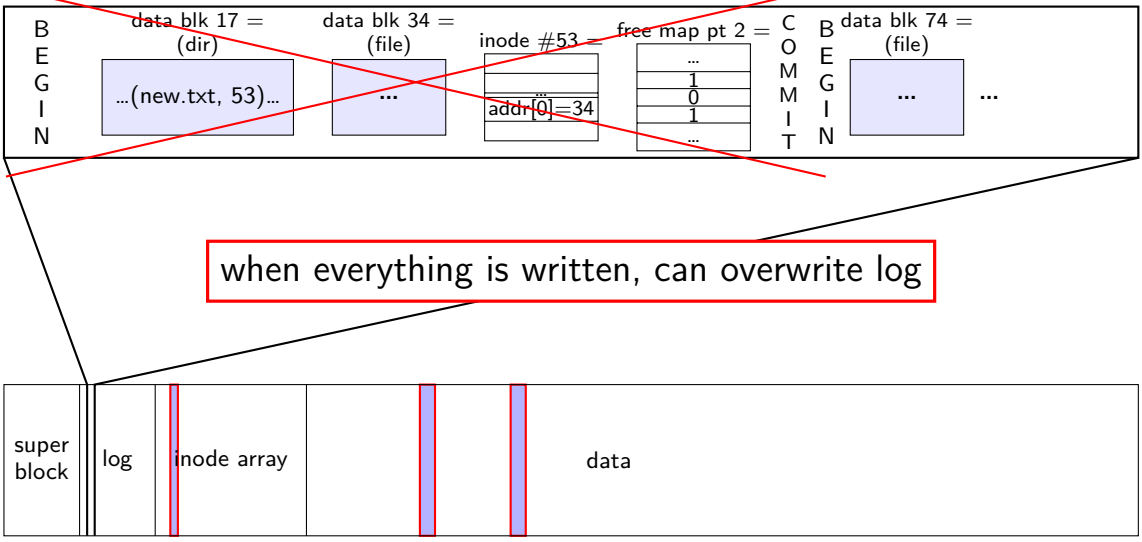
redo logging: file creation



redo logging: file creation



redo logging: file creation



redo logging: file creation

normal operation

write to log transaction steps:

- data blocks to create
- direcotry entry, inode to write
- directory inode (size, time)
- update

write to log "commit transaction"

in any order:

- update file data blocks
- update directory entry
- update file inode
- update directory inode

reclaim space in log

- "garbage collection"

redo logging: file creation

normal operation

write to log transaction steps:

- data blocks to create
- direcotry entry, inode to write
- directory inode (size, time)
- update

write to log "commit transaction"

in any order:

- update file data blocks
- update directory entry
- update file inode
- update directory inode

reclaim space in log

- "garbage collection"

crash before *commit*?
file not created
no partial operation to real data

redo logging: file creation

normal operation

write to log transaction steps:

- data blocks to create
- direcotry entry, inode to write
- directory inode (size, time)
- update

write to log "commit transaction"

in any order:

- update file data blocks
- update directory entry
- update file inode
- update directory inode

reclaim space in log

- "garbage collection"

crash after *commit*?

file created

promise: will perform logged updates

(after system reboots/recovers)

redo logging: file creation

normal operation

write to log transaction steps:

- data blocks to create
- direcotry entry, inode to write
- directory inode (size, time)
- update

write to log "commit transaction"

in any order:

- update file data blocks
- update directory entry
- update file inode
- update directory inode

reclaim space in log

- "garbage collection"

redo logging: file creation

normal operation

write to log transaction steps:

- data blocks to create
- direcotry entry, inode to write
- directory inode (size, time) update

write to log "commit transaction"

in any order:

- update file data blocks
- update directory entry
- update file inode
- update directory inode

reclaim space in log

- "garbage collection"

recovery

read log and...

ignore any operation with no "commit"

redo any operation with "commit"

- already done? — okay, setting inode twice

reclaim space in log

idempotency

logged operations should be *okay to do twice* = *idempotent*

good example: set inode link count to 4

bad example: increment inode link count

good example: overwrite inode number X with new value

as long as last committed inode value in log is right...

bad example: allocate new inode with particular contents

good example: overwrite data block with new value

bad example: append data to last used block of file

redo logging summary

write intended operation to the log

- before ever touching 'real' data
- in format that's safe to do twice

write marker to commit to the log

- if exists, the operation *will be done eventually*

actually update the real data

redo logging and filesystems

filesystems that do redo logging are called *journalling filesystems*

exercise (1)

suppose OS performing operation of appending 100KB to a 100KB file X in directory Y and uses redo logging, ext2-like filesystem with 1KB blocks, 4B block pointers

part 1: what's modified?

- [A] free block map
- [B] data blocks for file
- [C] indirect blocks for file
- [D] data blocks for directory
- [E] inode for file
- [F] inode for directory
- [G] the log

exercise (2)

suppose OS performing operation of appending 100KB to a 100KB file X in directory Y and uses redo logging

part 2: crash happens after writing:

- log entries for entire operation
- free block map changes
- indirect blocks for file

...what is written after restart as part of this operation?

- [A] free block map
- [B] data blocks for file
- [C] indirect blocks for file
- [D] data blocks for directory
- [E] inode for file
- [F] inode for directory
- [G] the log

lots of writing?

entire log can be written sequentially

- ideal for hard disk performance
- also pretty good for SSDs

no waiting for 'real' updates

- application can proceed while updates are happening
- files will be updated even if system crashes

often better for performance!

degrees of consistency

not all journalling filesystem use redo logging for everything

some use it *only for metadata operations*

some use it *for both metadata and user data*

only metadata: avoids lots of duplicate writing

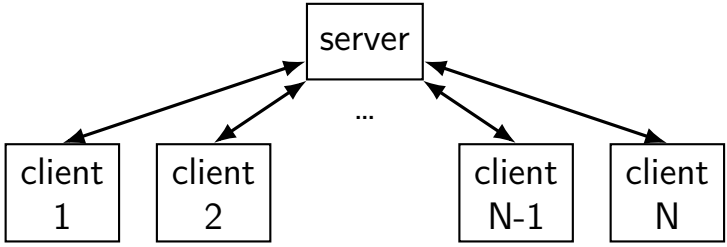
metadata+user data: integrity of user data guaranteed

distributed systems

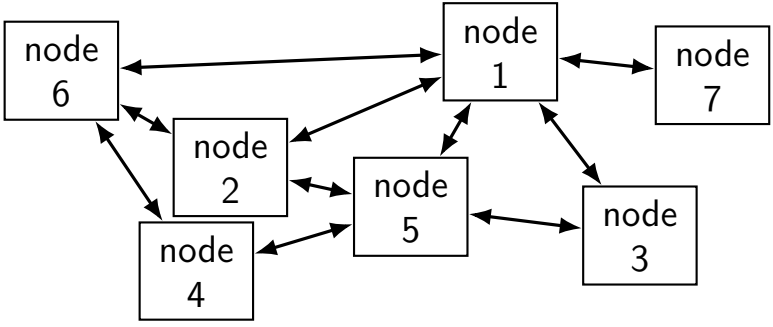
multiple machines working together to perform a single task

called a *distributed system*

some distibuted systems models

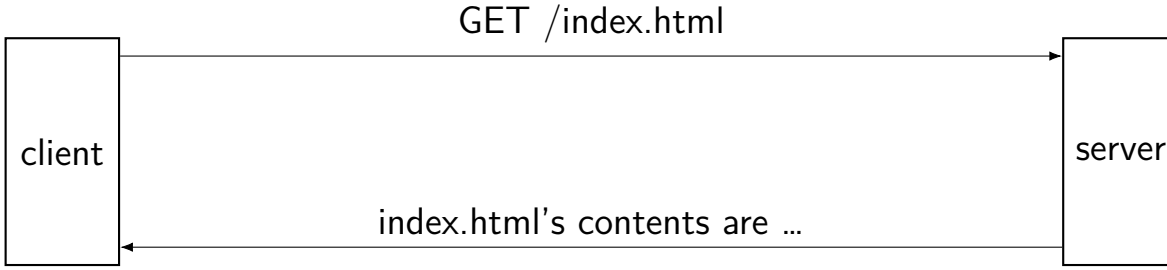


client/server

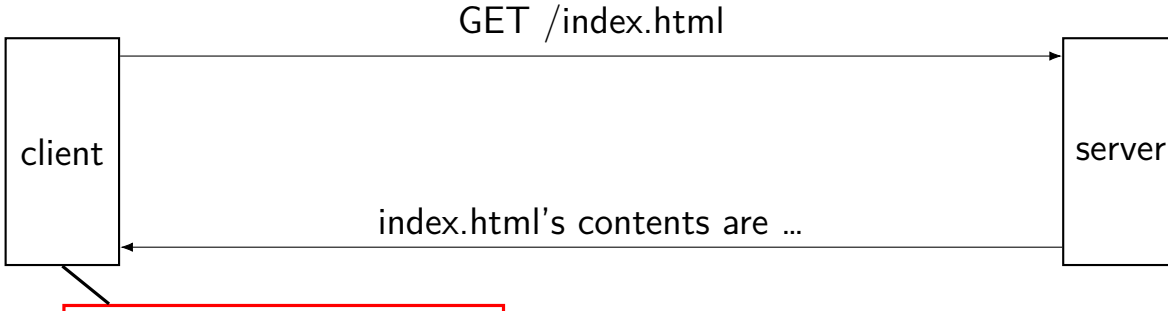


peer-to-peer

client/server model



client/server model

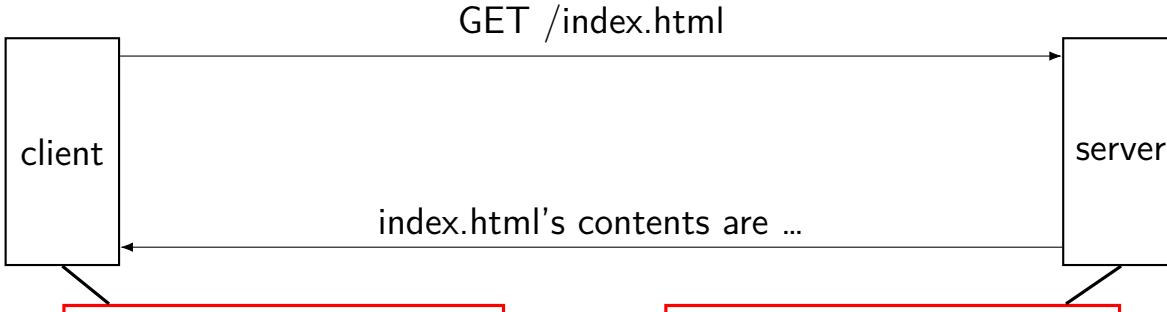


client(s): "sometimes on"

sends requests to server(s)

needs to know
how to contact server

client/server model



client(s): "sometimes on"

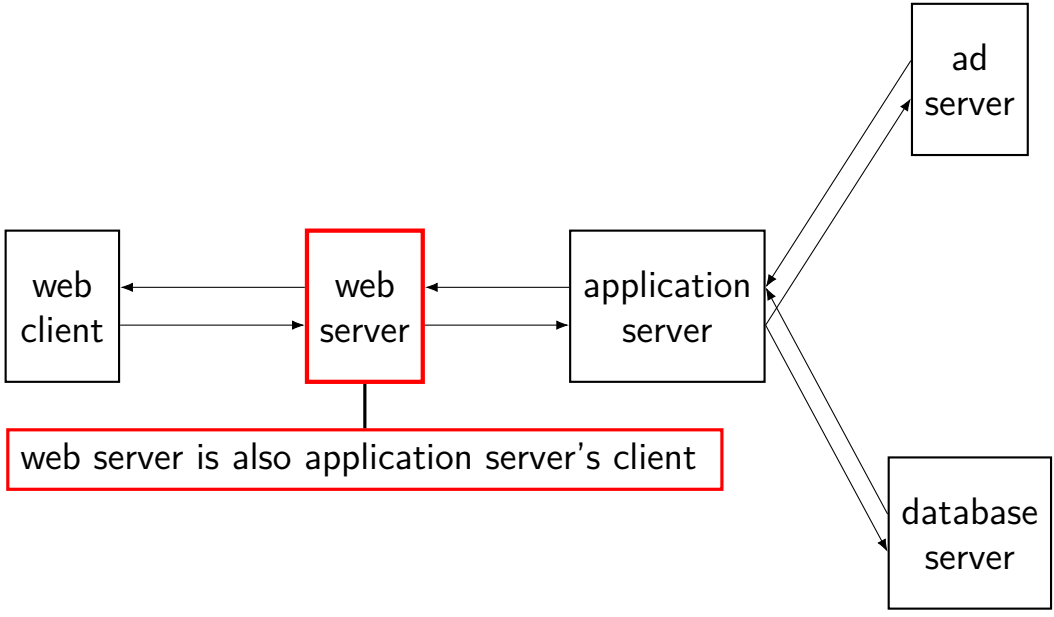
sends requests to server(s)

needs to know
how to contact server

server(s): "always on"

responds to client requests
never initiaties contact
with a client

layers of servers?



example: Wikipedia architecture

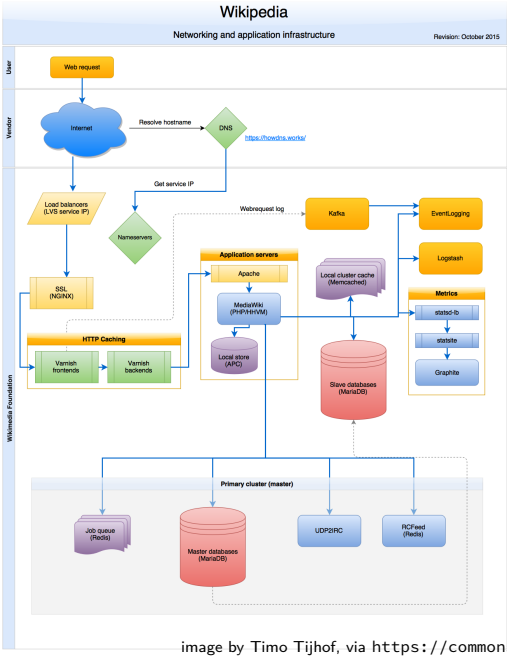


image by Timo Tijhof, via https://commons.wikimedia.org/wiki/File:Wikipedia_webrequest_flow_2015-10.png

example: Wikipedia architecture (zoom)

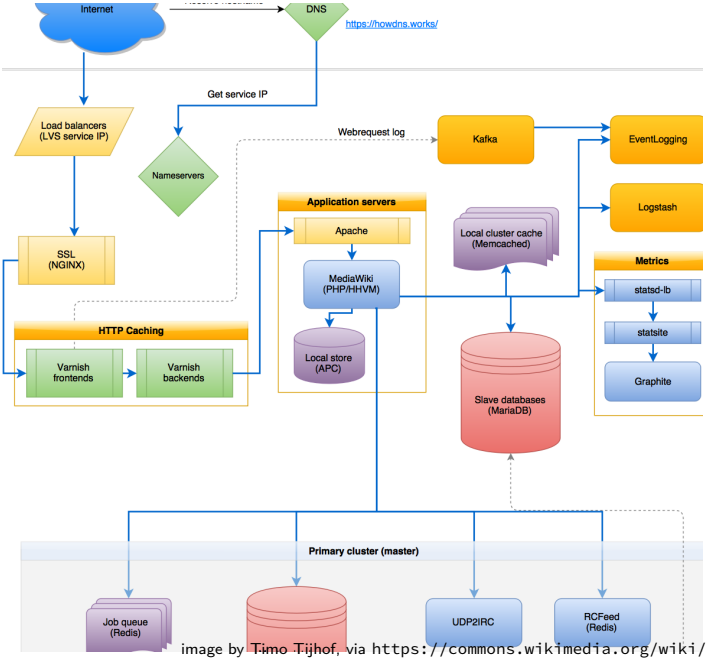


image by Timo Tijhof, via https://commons.wikimedia.org/wiki/File:Wikipedia_webrequest_flow_2015-10.png

peer-to-peer

no always-on server everyone knows about

- hopefully, no one bottleneck — "scalability"

any machine can contact any other machine

- every machine plays an approx. equal role?

set of machines may change over time

why distributed?

multiple machine owners collaborating

delegation of responsiblity to other entity

put (part of) service "in the cloud"

combine many cheap machines to replace expensive machine

easier to add incrementally

redundancy — one machine can fail and *system* still works?

exercise

which are likely advantages of client/server model over peer-to-peer?

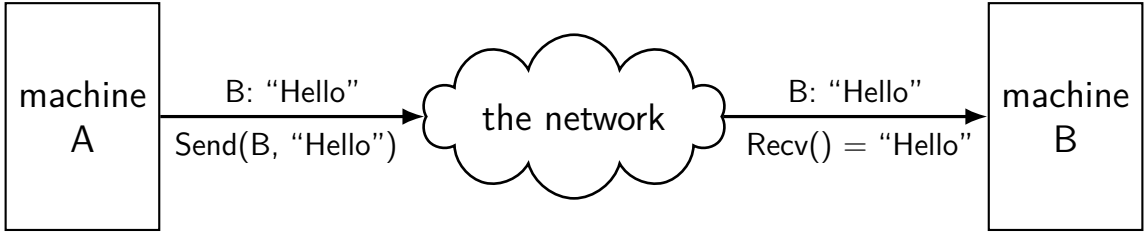
[A] easier to make whole system work despite failure of any machine

[B] easier to handle most machines being offline a majority of the time

[C] better suited to a mix of a few very big/high-performance and many small/low-performance machines

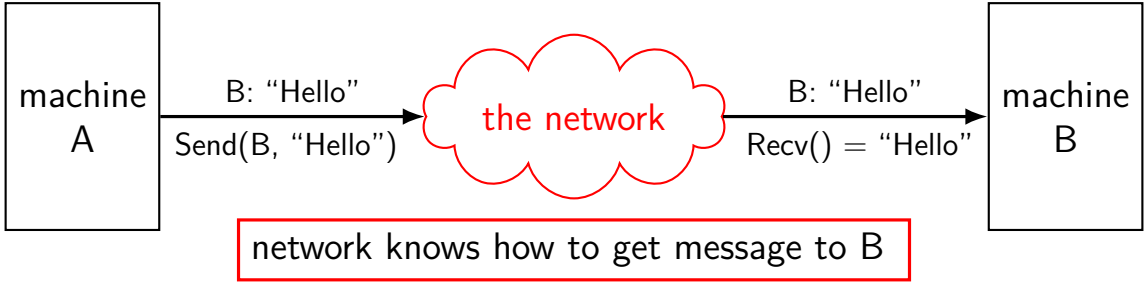
mailbox model

mailbox abstraction: send/receive messages



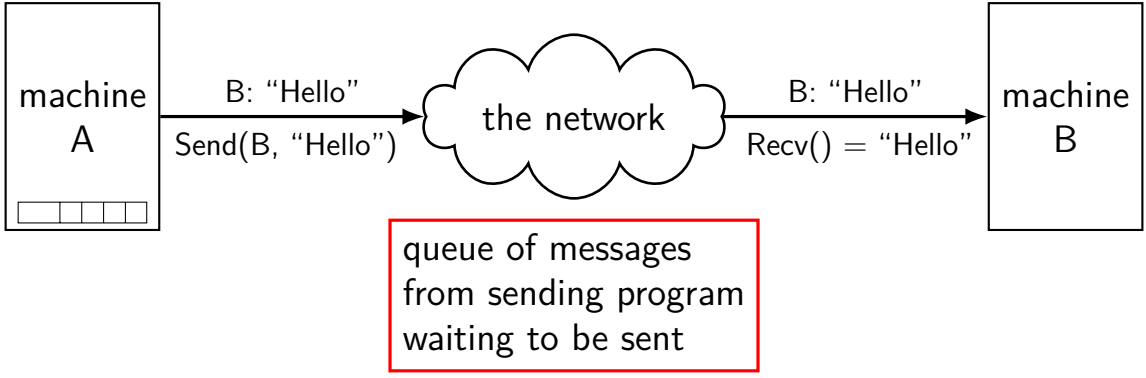
mailbox model

mailbox abstraction: send/receive messages



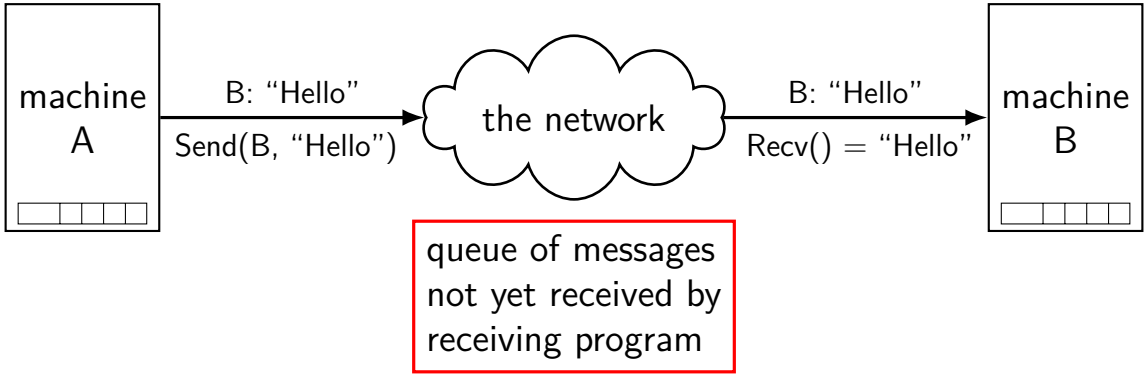
mailbox model

mailbox abstraction: send/receive messages



mailbox model

mailbox abstraction: send/receive messages



what about servers?

client/server model: server wants to reply to clients

might want to send/receive multiple messages

what about servers?

client/server model: server wants to reply to clients

might want to send/receive multiple messages

can build this with mailbox idea

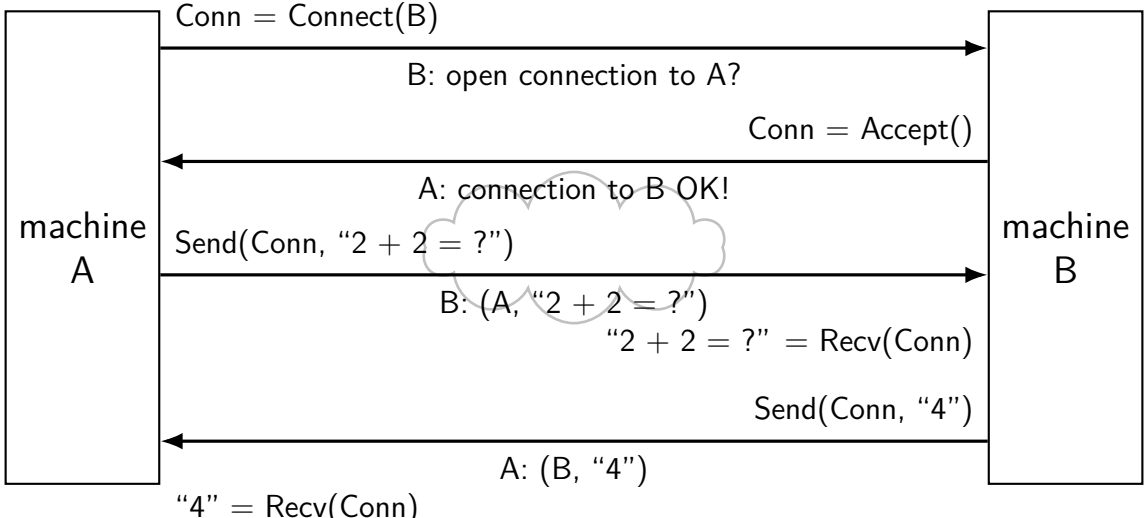
- send a 'return address'
- need to track related messages

common abstraction that does this: the connection

extension: conections

connections: two-way channel for messages

extra operations: connect, accept



connections versus pipes

connections look kinda like two-direction pipes

in fact, in POSIX will have the same API:

each end gets file descriptor representing connection

can use read() and write()

connections over mailboxes

real Internet: mailbox-style communication

- send packets to particular mailboxes
- no gaurentee on order, when received
- no relationship between

connections implemented on top of this

full details: take networking (CS/ECE 4457)

connection missing pieces?

how to specify the machine?

multiple programs on one machine? who gets the message?

names and addresses

name	**address**
logical identifier	location/how to locate
hostname `www.virginia.edu`	IPv4 address `128.143.22.36`
hostname `mail.google.com`	IPv4 address `216.58.217.69`
hostname `mail.google.com`	IPv6 address `2607:f8b0:4004:80b::2005`
filename `/home/cr4bd/NOTES.txt`	inode# `120800873` and device `0x2eh/0x46d`
variable `counter`	memory address `0x7FFF9430`
service name `https`	port number 443

hostnames

typically use *domain name system* (DNS) to find machine names

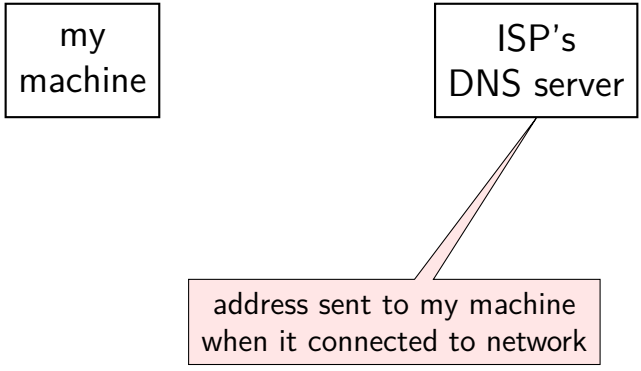
maps logical names like `www.virginia.edu`

- chosen for humans
- hierarchy of names

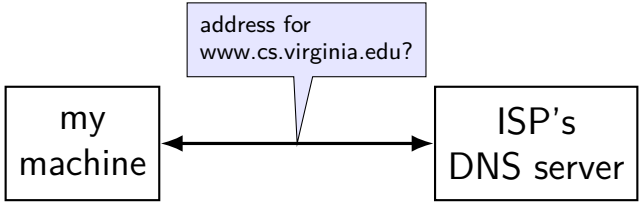
...to *addresses* the network can use to move messages

- numbers
- ranges of numbers assigned to different parts of the network
- network *routers* knows "send this range of numbers goes this way"

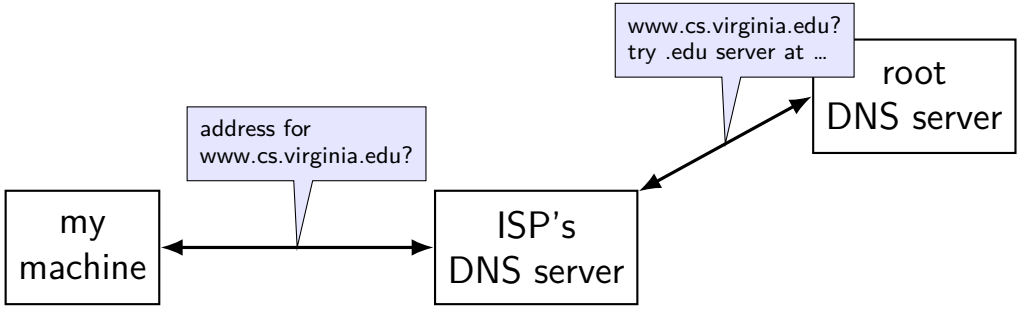
DNS: distributed database



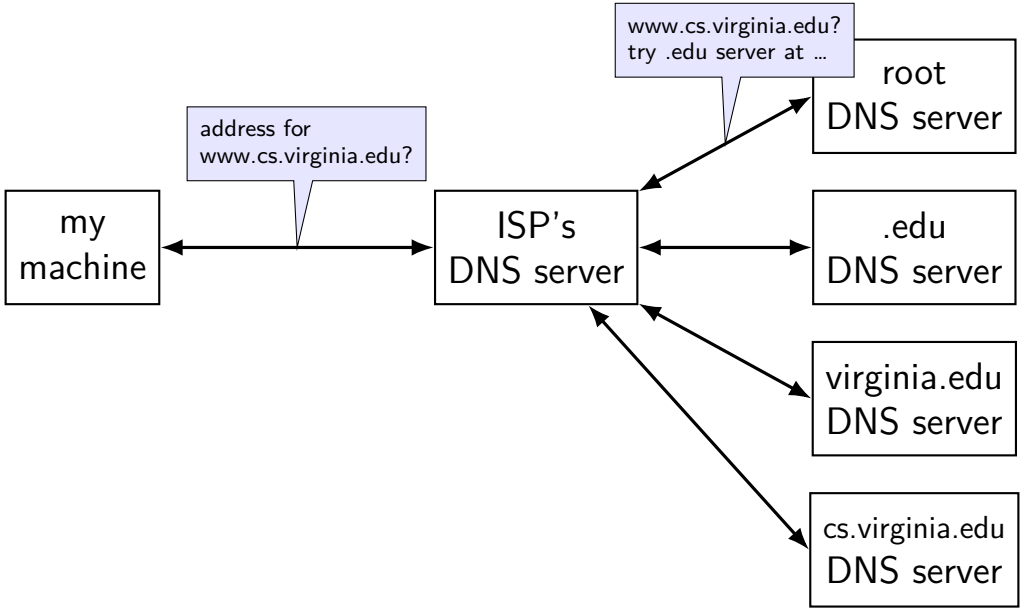
DNS: distributed database



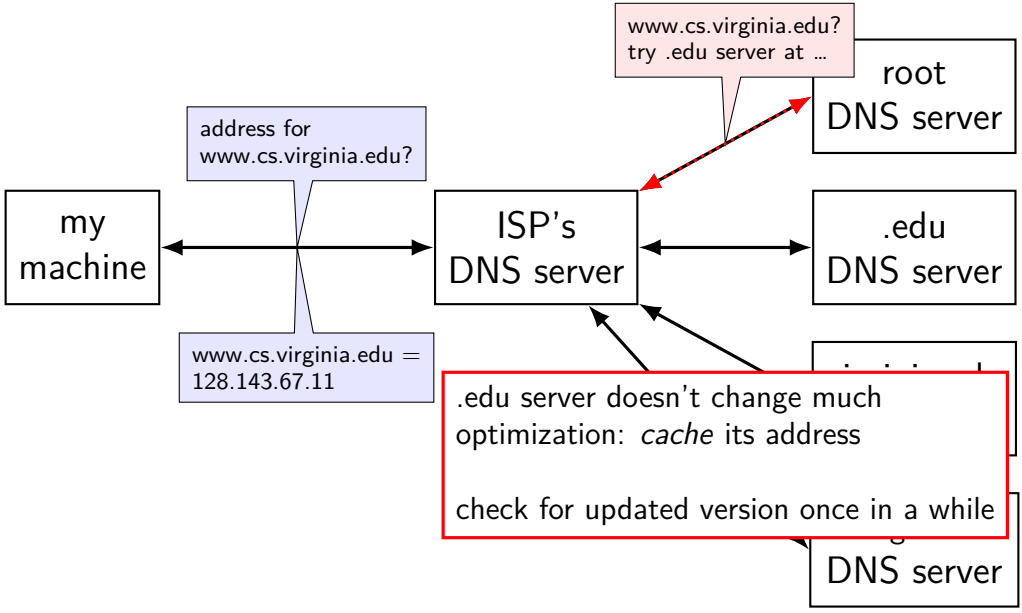
DNS: distributed database



DNS: distributed database



DNS: distributed database



connection missing pieces?

how to specify the machine?

multiple programs on one machine? who gets the message?

IPv4 addresses

32-bit numbers

typically written like 128.143.67.11

- four 8-bit decimal values separated by dots
- first part is most significant
- same as $128 \cdot 256^3 + 143 \cdot 256^2 + 67 \cdot 256 + 11 = 2\,156\,782\,459$

organizations get blocks of IPs

- e.g. UVa has 128.143.0.0–128.143.255.255
- e.g. Google has 216.58.192.0–216.58.223.255 and 74.125.0.0–74.125.255.255 and 35.192.0.0–35.207.255.255

selected special IPv4 addresses

127.0.0.0 — 127.255.255.255 — localhost

- AKA loopback
- the machine we're on
- typically only 127.0.0.1 is used

192.168.0.0–192.168.255.255 and
10.0.0.0–10.255.255.255 and
172.16.0.0–172.31.255.255

- "private" IP addresses
- not used on the Internet
- commonly connected to Internet with network address translation
- also 100.64.0.0–100.127.255.255 (but with restrictions)

169.254.0.0-169.254.255.255

- link-local addresses — 'never' forwarded by routers

network address translation

IPv4 addresses are kinda scarce

solution: *convert* many private addrs. to one public addr.

locally: use private IP addresses for machines

outside: private IP addresses become a single public one

commonly how home networks work (and some ISPs)

IPv6 addresses

IPv6 like IPv4, but with 128-bit numbers

written in hex, 16-bit parts, seperated by colons (`:`)

strings of 0s represented by double-colons (`::`)

typically given to users in blocks of 2^{80} or 2^{64} addresses

- no need for address translation?

`2607:f8b0:400d:c00::6a` =
`2607:f8b0:400d:0c00:0000:0000:0000:006a`

- `2607f8b0400d0c0000000000000006a`$_{\text{SIXTEEN}}$

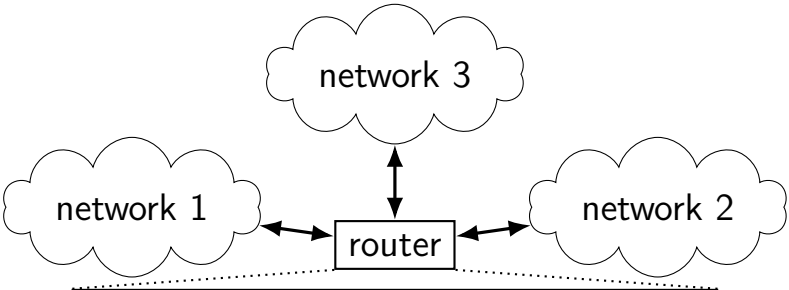
selected special IPv6 addresses

`::1` = localhost

anything starting with `fe80` = link-local addresses

- never forwarded by routers

IPv4 addresses and routing tables



if I receive data for...	send it to...
128.143.0.0—128.143.255.255	network 1
192.107.102.0–192.107.102.255	network 1
...	...
4.0.0.0–7.255.255.255	network 2
64.8.0.0–64.15.255.255	network 2
...	...
anything else	network 3

connection missing pieces?

how to specify the machine?

multiple programs on one machine? who gets the message?

port numbers

we run multiple programs on a machine

- IP addresses identifying machine — not enough

port numbers

we run multiple programs on a machine

- IP addresses identifying machine — not enough

so, add 16-bit *port numbers*

- think: multiple PO boxes at address

port numbers

we run multiple programs on a machine

- IP addresses identifying machine — not enough

so, add 16-bit *port numbers*

- think: multiple PO boxes at address

0–49151: typically assigned for particular services

- 80 = http, 443 = https, 22 = ssh, ...

49152–65535: allocated on demand

- default "return address" for client connecting to server

protocols

protocol = agreement on how to comunicate

syntax (format of messages, etc.)

- e.g. mailbox model: where does address go?
- e.g. connection: where does return address go?

semantics (meaning of messages — actions to take, etc.)

- e.g. connection: when to consider connection created?

human protocol: telephone

caller: pick up phone caller: check for service caller: dial caller: wait for ringing	
	callee: “Hello?”
caller: “Hi, it’s Casey...”	
	callee: “Hi, so how about ...”
caller: “Sure, ...”	
...	...
	callee: “Bye!”
caller: “Bye!”	
hang up	hang up

layered protocols

IP: protocol for sending data by IP addresses

- mailbox model
- limited message size

UDP: send *datagrams* built on IP

- still mailbox model, but *with port numbers*

TCP: reliable connections built on IP

- adds port numbers
- adds resending data if error occurs
- splits big amounts of data into many messages

HTTP: protocol for sending files, etc. built on TCP

other notable protocols (transport layer)

TLS: Transport Layer Security — built on TCP

like TCP, but adds encryption + authentication

SSH: secure shell (remote login) — built on TCP

SCP/SFTP: secure copy/secure file transfer — built on SSH

HTTPS: HTTP, but over TLS instead of TCP

FTP: file transfer protocol

...

sockets

socket: POSIX abstraction of network I/O queue

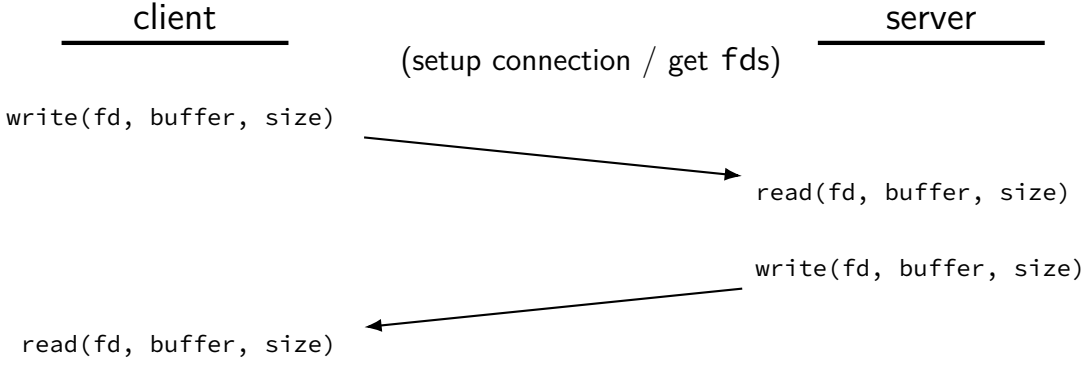
- any kind of network
- can also be used between processes on same machine

a kind of file descriptor

connected sockets

sockets can represent a connection

act like bidirectional pipe



echo client/server

```
void client_for_connection(int socket_fd) {
    int n; char send_buf[MAX_SIZE]; char recv_buf[MAX_SIZE];
    while (prompt_for_input(send_buf, MAX_SIZE)) {
        n = write(socket_fd, send_buf, strlen(send_buf));
        if (n != strlen(send_buf)) {...error?...}
        n = read(socket_fd, recv_buf, MAX_SIZE);
        if (n <= 0) return; // error or EOF
        write(STDOUT_FILENO, recv_buf, n);
    }
}
```

```
void server_for_connection(int socket_fd) {
    int read_count, write_count; char request_buf[MAX_SIZE];
    while (1) {
        read_count = read(socket_fd, request_buf, MAX_SIZE);
        if (read_count <= 0) return; // error or EOF
        write_count = write(socket_fd, request_buf, read_count);
        if (read_count != write_count) {...error?...}
    }
}
```

echo client/server

```
void client_for_connection(int socket_fd) {
    int n; char send_buf[MAX_SIZE]; char recv_buf[MAX_SIZE];
    while (prompt_for_input(send_buf, MAX_SIZE)) {
        n = write(socket_fd, send_buf, strlen(send_buf));
        if (n != strlen(send_buf)) {...error?...}
        n = read(socket_fd, recv_buf, MAX_SIZE);
        if (n <= 0) return; // error or EOF
        write(STDOUT_FILENO, recv_buf, n);
    }
}
```

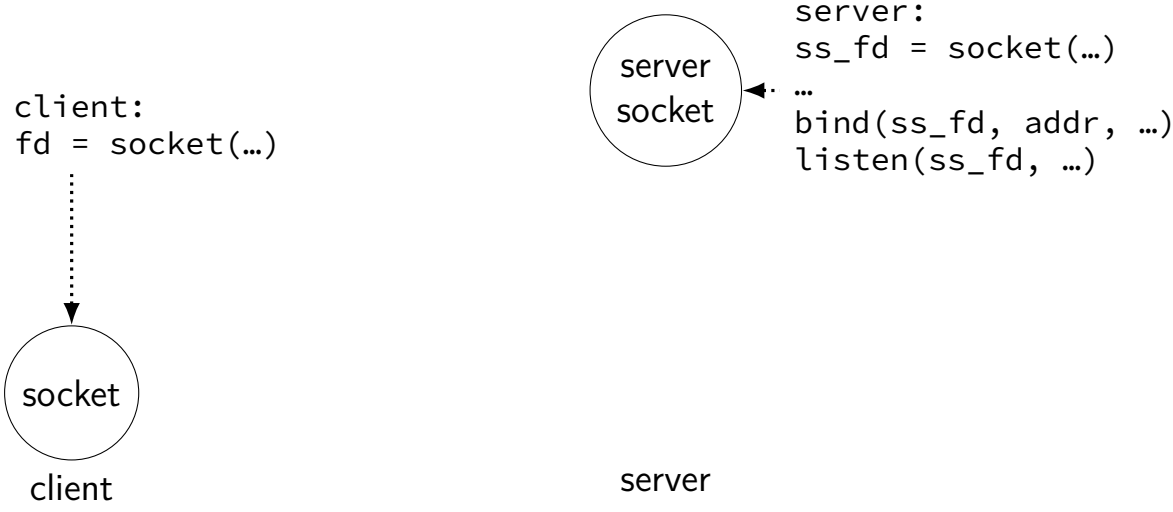
```
void server_for_connection(int socket_fd) {
    int read_count, write_count; char request_buf[MAX_SIZE];
    while (1) {
        read_count = read(socket_fd, request_buf, MAX_SIZE);
        if (read_count <= 0) return; // error or EOF
        write_count = write(socket_fd, request_buf, read_count);
        if (read_count != write_count) {...error?...}
    }
}
```

echo client/server

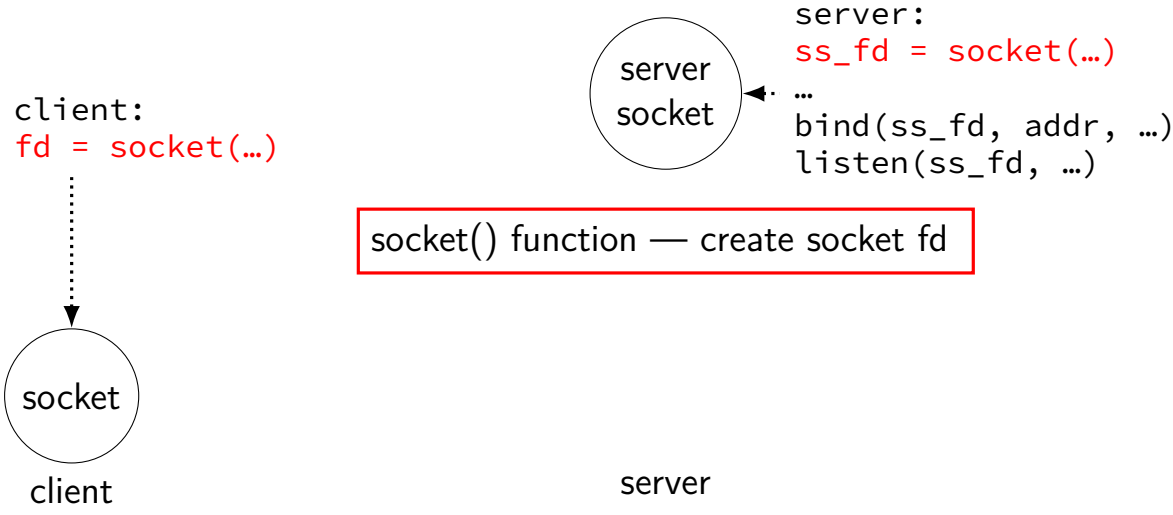
```
void client_for_connection(int socket_fd) {
    int n; char send_buf[MAX_SIZE]; char recv_buf[MAX_SIZE];
    while (prompt_for_input(send_buf, MAX_SIZE)) {
        n = write(socket_fd, send_buf, strlen(send_buf));
        if (n != strlen(send_buf)) {...error?...}
        n = read(socket_fd, recv_buf, MAX_SIZE);
        if (n <= 0) return; // error or EOF
        write(STDOUT_FILENO, recv_buf, n);
    }
}
```

```
void server_for_connection(int socket_fd) {
    int read_count, write_count; char request_buf[MAX_SIZE];
    while (1) {
        read_count = read(socket_fd, request_buf, MAX_SIZE);
        if (read_count <= 0) return; // error or EOF
        write_count = write(socket_fd, request_buf, read_count);
        if (read_count != write_count) {...error?...}
    }
}
```

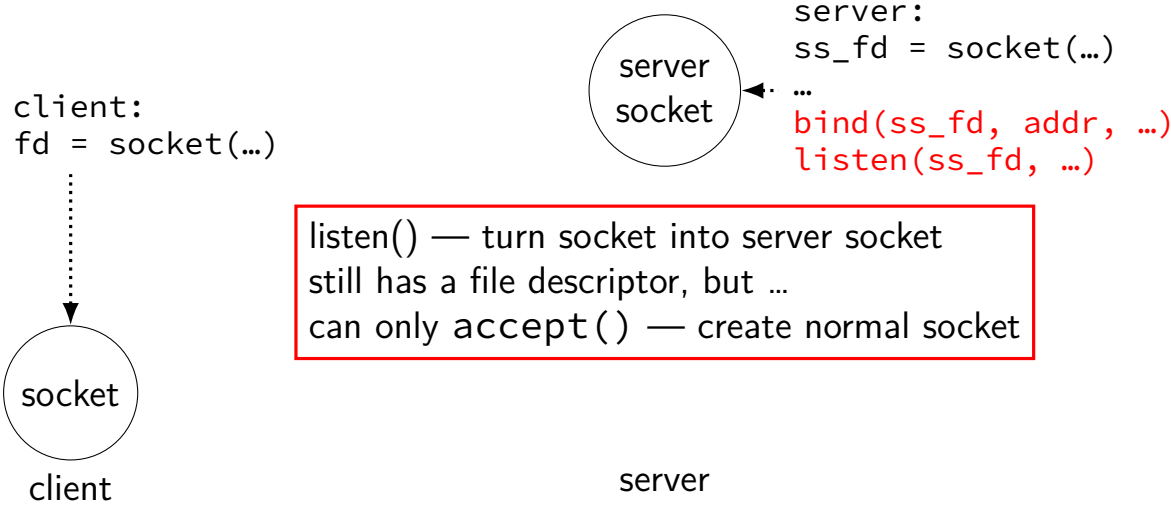
sockets and server sockets



sockets and server sockets

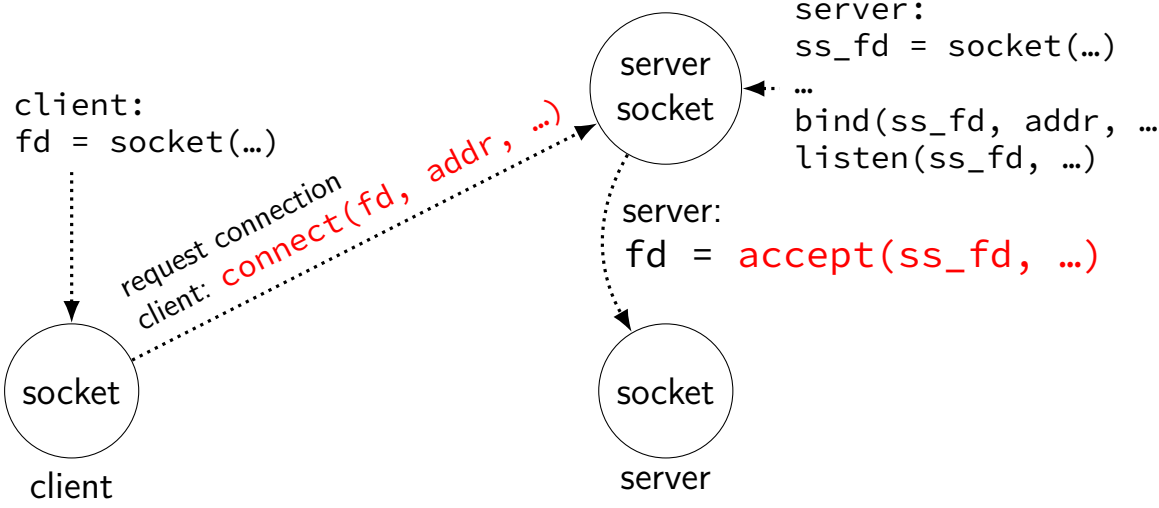


sockets and server sockets

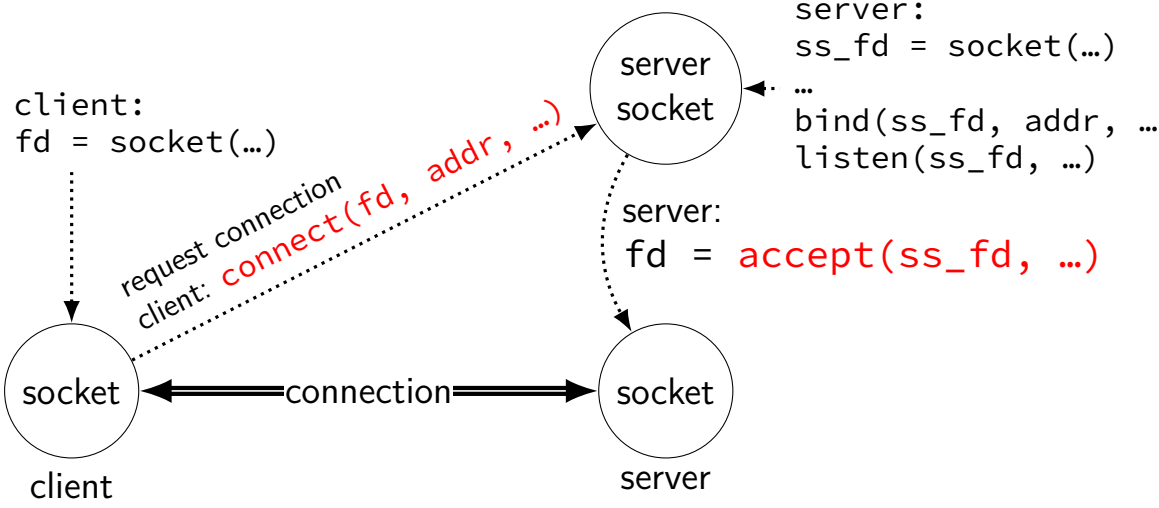


listen() — turn socket into server socket
still has a file descriptor, but ...
can only `accept()` — create normal socket

sockets and server sockets



sockets and server sockets



connections in TCP/IP

on network: connection identified by *5-tuple*

used by OS to lookup "where is the file descriptor?"

(protocol=TCP, local IP addr., local port, remote IP addr., remote port)

both ends always have an address+port

what is the IP address, port number? set with `bind()` function

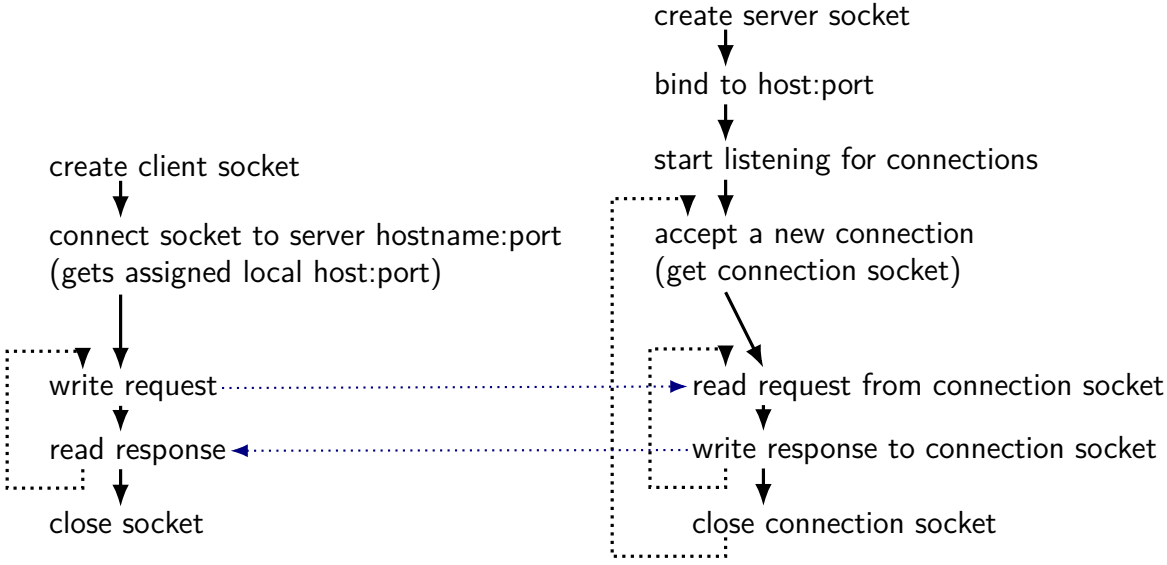
typically always done for servers, not done for clients

system will choose default if you don't

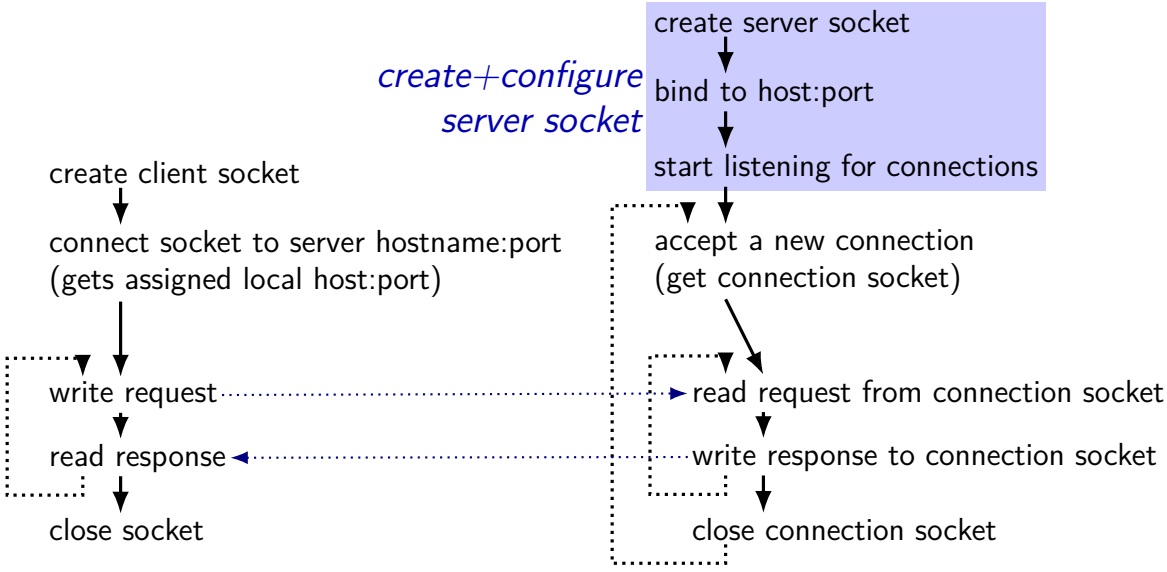
connections on my desktop

```
cr4bd@reiss-t3620
: /zf14/cr4bd ; netstat --inet --inet6 --numeric
Active Internet connections (w/o servers)
Proto Recv-Q Send-Q Local Address           Foreign Address         State
tcp        0      0 128.143.67.91:49202     128.143.63.34:22        ESTABLISHE
tcp        0      0 128.143.67.91:803       128.143.67.236:2049     ESTABLISHE
tcp        0      0 128.143.67.91:50292     128.143.67.226:22       TIME_WAIT
tcp        0      0 128.143.67.91:54722     128.143.67.236:2049     TIME_WAIT
tcp        0      0 128.143.67.91:52002     128.143.67.236:111      TIME_WAIT
tcp        0      0 128.143.67.91:732       128.143.67.236:63439    TIME_WAIT
tcp        0      0 128.143.67.91:40664     128.143.67.236:2049     TIME_WAIT
tcp        0      0 128.143.67.91:54098     128.143.67.236:111      TIME_WAIT
tcp        0      0 128.143.67.91:49302     128.143.67.236:63439    TIME_WAIT
tcp        0      0 128.143.67.91:50236     128.143.67.236:111      TIME_WAIT
tcp        0      0 128.143.67.91:22        172.27.98.20:49566      ESTABLISHE
tcp        0      0 128.143.67.91:51000     128.143.67.236:111      TIME_WAIT
tcp        0      0 127.0.0.1:50438         127.0.0.1:631           ESTABLISHE
tcp        0      0 127.0.0.1:631           127.0.0.1:50438         ESTABLISHE
```

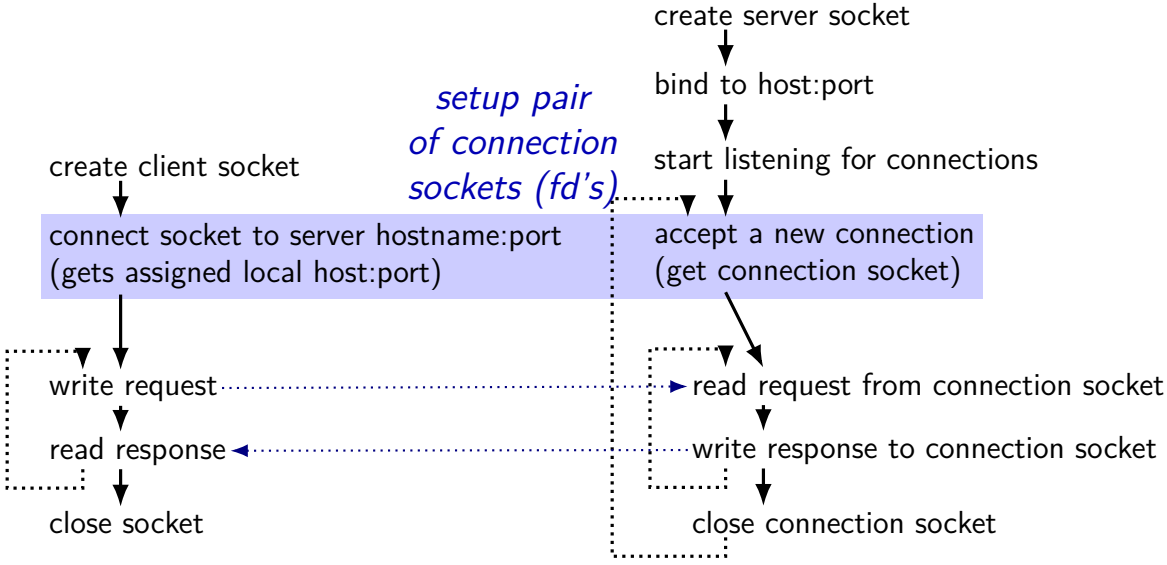
client/server flow (one connection at a time)



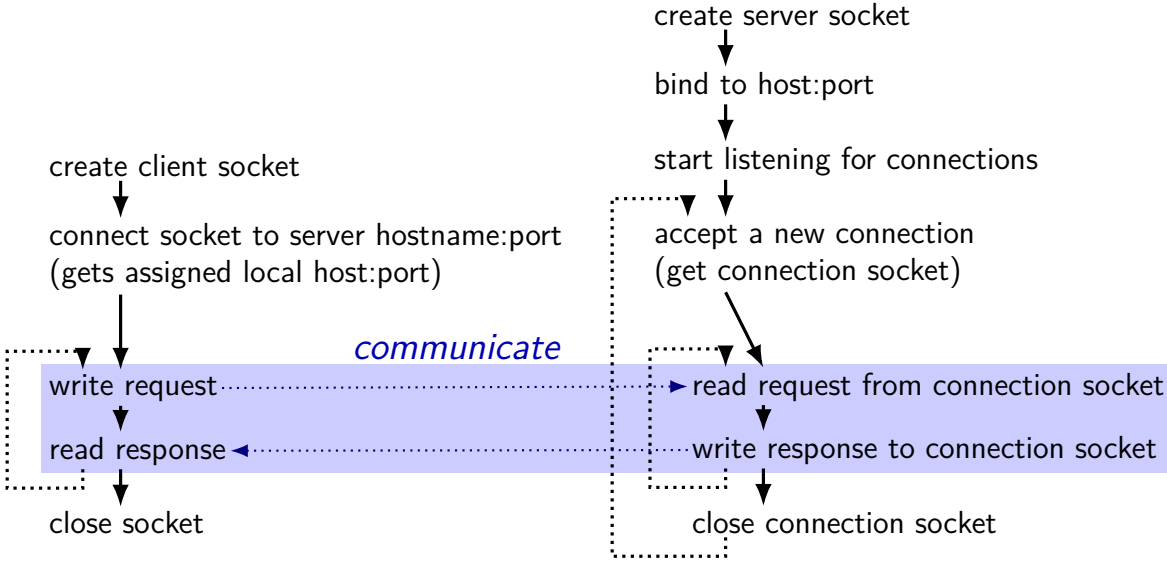
client/server flow (one connection at a time)



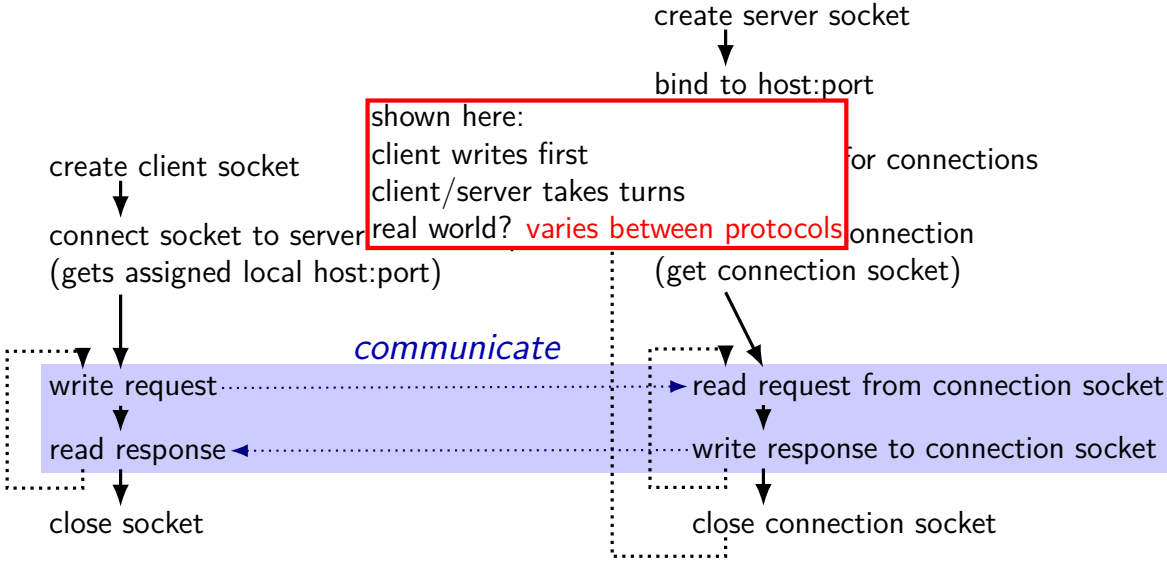
client/server flow (one connection at a time)



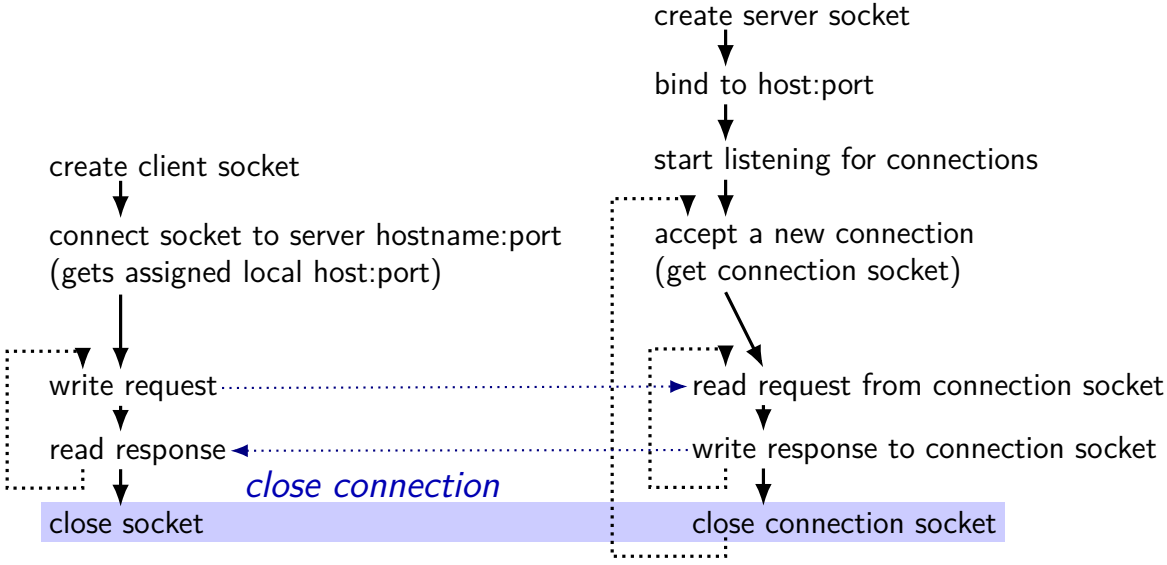
client/server flow (one connection at a time)



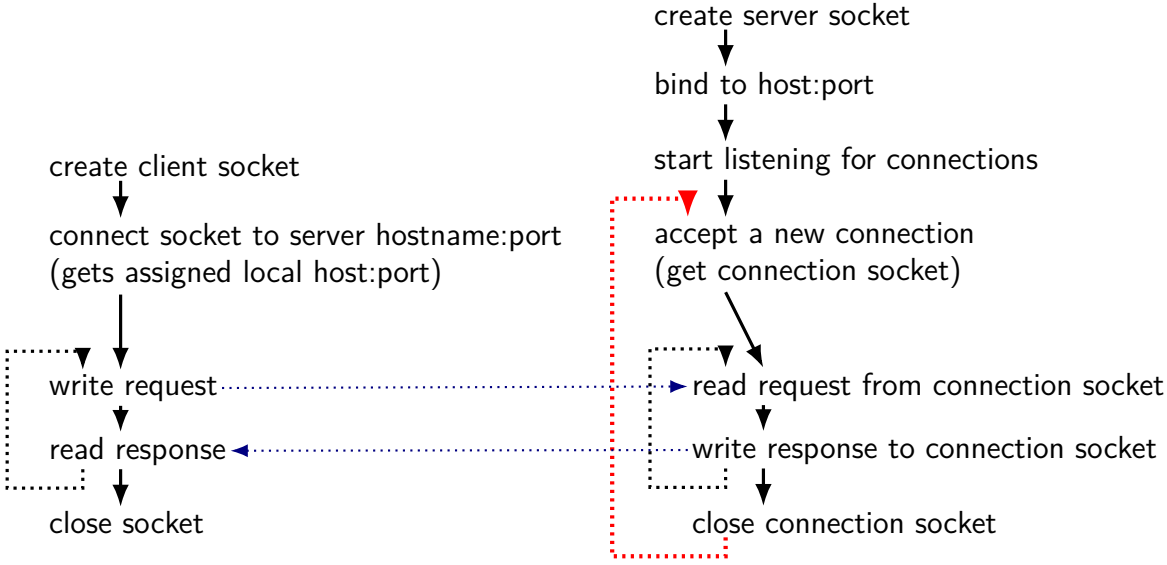
client/server flow (one connection at a time)



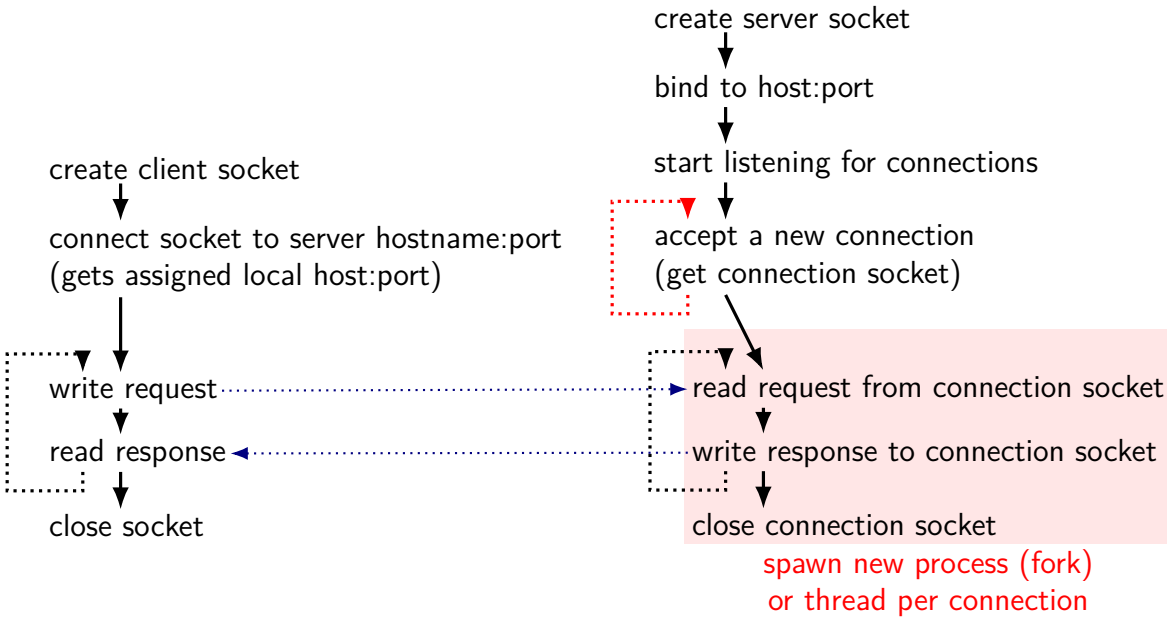
client/server flow (one connection at a time)



client/server flow (one connection at a time)

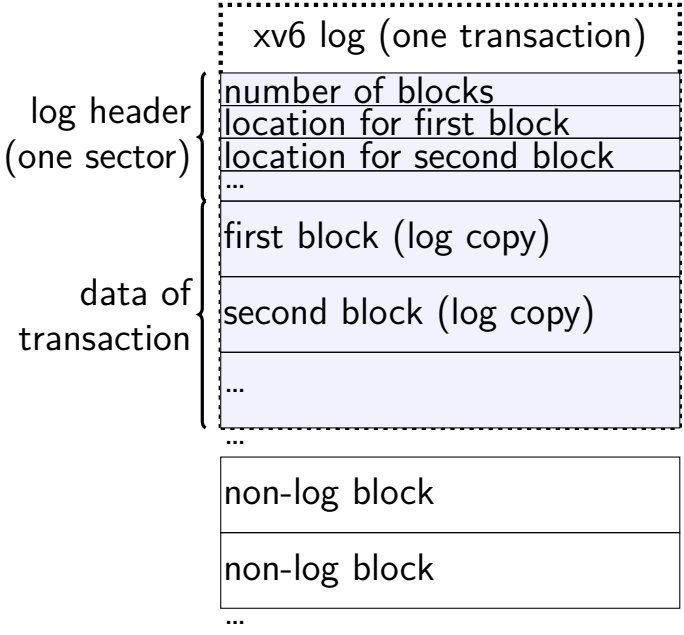


client/server flow (multiple connections)

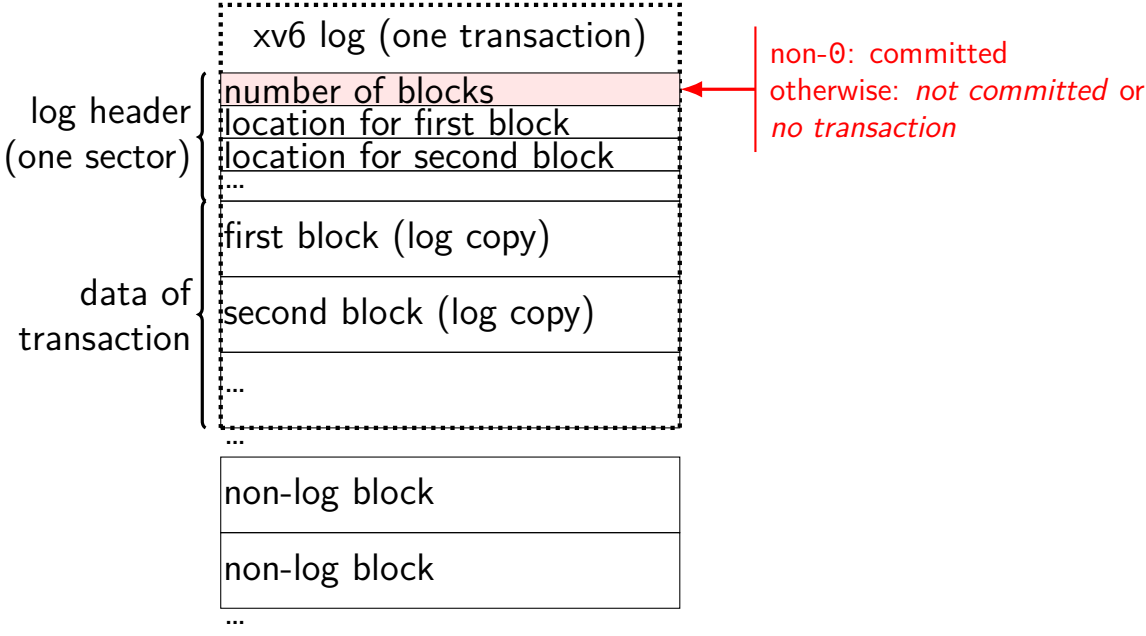


backup slides

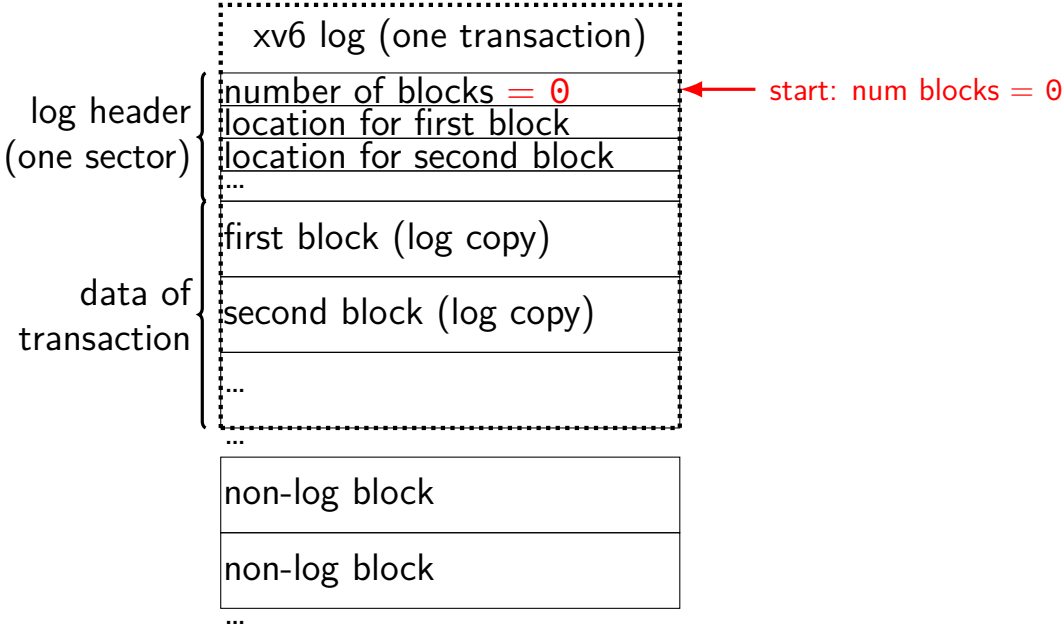
the xv6 journal



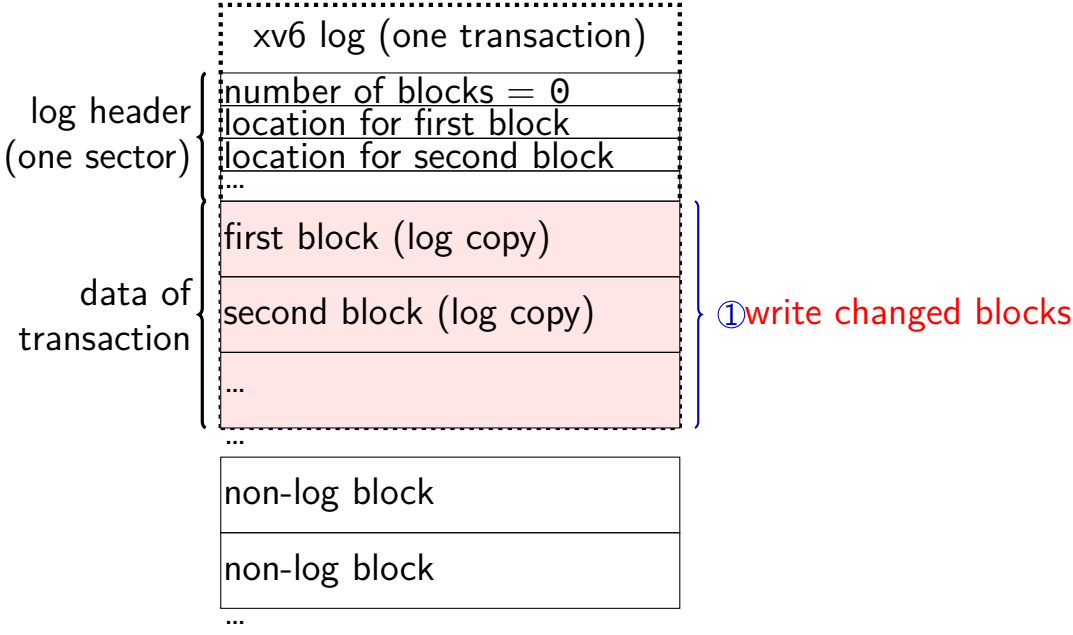
the xv6 journal



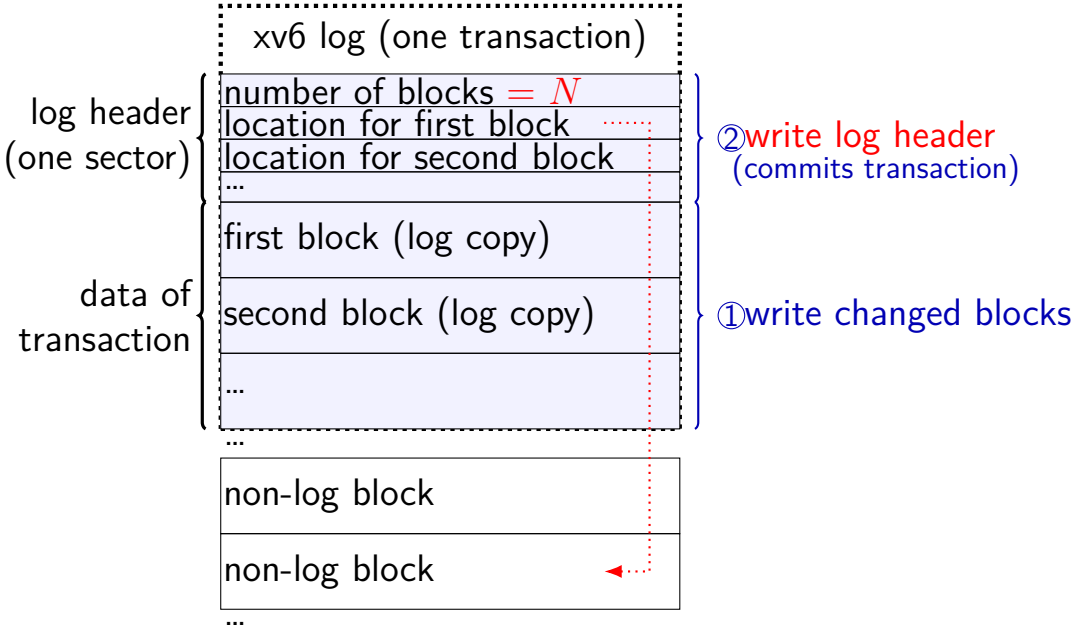
the xv6 journal



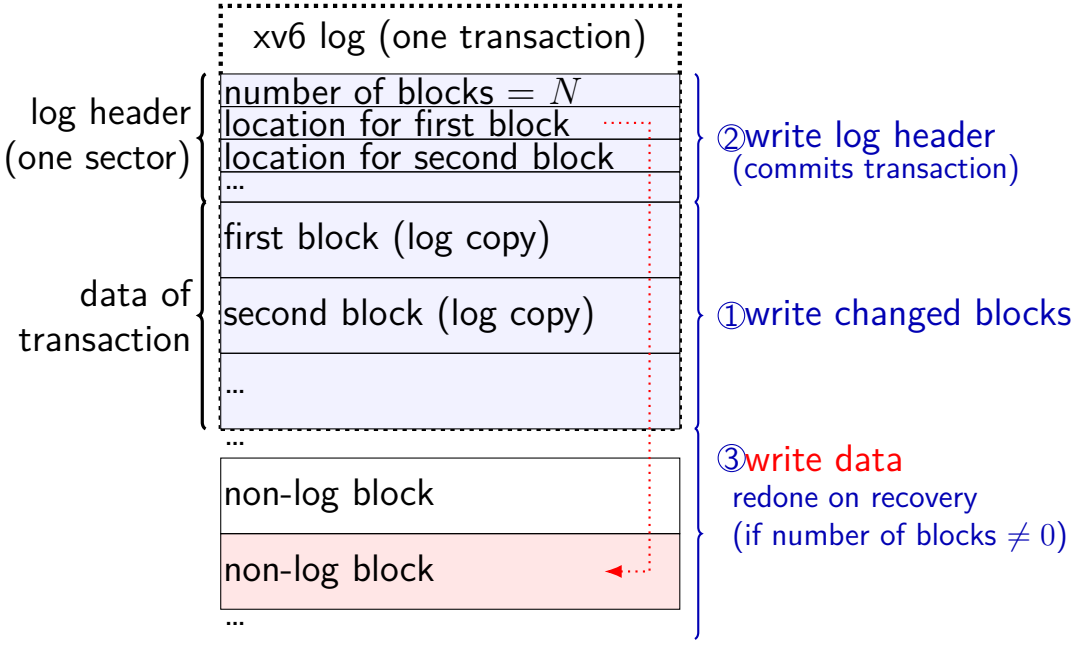
the xv6 journal



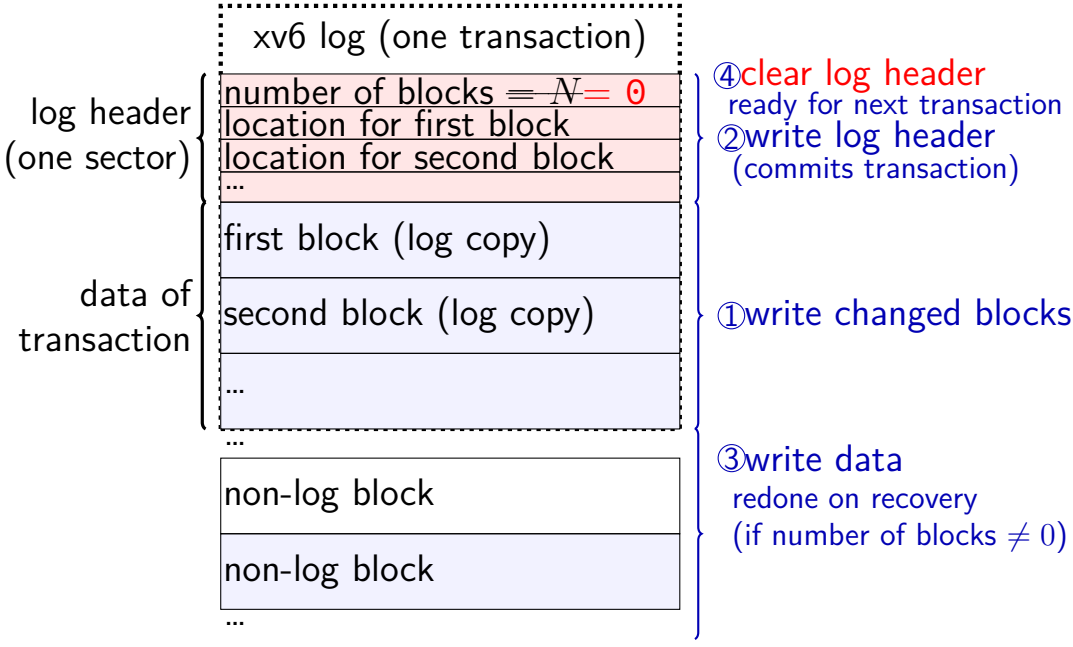
the xv6 journal



the xv6 journal



the xv6 journal



what is a transaction?

so far: each file update?

faster to do batch of updates together

- one log write finishes lots of things
- don't wait to write

xv6 solution: combine lots of updates into one transaction

only commit when...

- no active file operation, *or*
- not enough room left in log for more operations

what is a transaction?

so far: each file update?

faster to do batch of updates together

- one log write finishes lots of things
- don't wait to write

xv6 solution: combine lots of updates into one transaction

only commit when...

- no active file operation, *or*
- not enough room left in log for more operations

redo logging problems

doesn't the log get infinitely big?

writing everything twice?

redo logging problems

doesn't the log get infinitely big?

writing everything twice?

limiting log size

once transaction is written to real data, can discard

sometimes called "garbage collecting" the log

may sometimes need to block to free up log space

- perform logged updates before adding more to log

hope: usually log cleanup happens "in the background"

redo logging problems

doesn't the log get infinitely big?

writing everything twice?

reading and writing at once

so far assumption: alternate between reading+writing

- sufficient for FTP assignment
- how many protocols work

"half-duplex"

don't have to use sockets this way, but tricky

threads: one reading thread, one writing thread *OR*

event-loop: use *non-blocking I/O* and select()/poll()/etc. functions

- non-blocking I/O setup with fcntl() function
- non-blocking write() fills up buffer as much as possible, then returns
- non-blocking read() returns what's in buffer, never waits for more

mounting filesystems

Unix-like system

root filesystem appears as /

other filesystems *appear as directory*

e.g. lab machines: my home dir is in filesystem at /net/zf15

directories that are filesystems look like normal directories

`/net/zf15/..` is `/net` (even though in different filesystems)

mounts on a dept. machine

```
/dev/sda1 on / type ext4 (rw,errors=remount-ro)
proc on /proc type proc (rw,noexec,nosuid,nodev)
...
udev on /dev type devtmpfs (rw,mode=0755)
devpts on /dev/pts type devpts (rw,noexec,nosuid,gid=5,mode=0620)
tmpfs on /run type tmpfs (rw,noexec,nosuid,size=10%,mode=0755)
...
/dev/sda3 on /localtmp type ext4 (rw)
...
zfs1:/zf2 on /net/zf2 type nfs (rw,hard,intr,proto=udp,nfsvers=3,
                                noacl,sloppy,addr=128.143.136.9)
zfs3:/zf19 on /net/zf19 type nfs (rw,hard,intr,proto=udp,nfsvers=3,
                                  noacl,sloppy,addr=128.143.67.236)
zfs4:/sw on /net/sw type nfs (rw,hard,intr,proto=udp,nfsvers=3,
                               noacl,sloppy,addr=128.143.136.9)
zfs3:/zf14 on /net/zf14 type nfs (rw,hard,intr,proto=udp,nfsvers=3,
                                  noacl,sloppy,addr=128.143.67.236)
...
```

kernel FS abstractions

Linux: *virtual file system* API

object-oriented, based on FFS-style filesystem

to implement a filesystem, create object types for:

- superblock (represents "header")
- inode (represents file)
- dentry (represents cached directory entry)
- file (represents *open file*)

common code handles directory traversal

- and caches directory traversals

common code handles file descriptors, etc.

connection setup: client, using addrinfo

```
int sock_fd;
struct addrinfo *server = /* code on next slide */;

sock_fd = socket(
    server->ai_family,
     // ai_family = AF_INET (IPv4) or AF_INET6 (IPv6) or ...
    server->ai_socktype,
     // ai_socktype = SOCK_STREAM (bytes) or ...
    server->ai_prototcol
     // ai_protocol = IPPROTO_TCP or ...
);
if (sock_fd < 0) { /* handle error */ }
if (connect(sock_fd, server->ai_addr, server->ai_addrlen) < 0) {
    /* handle error */
}
freeaddrinfo(server);
DoClientStuff(sock_fd); /* read and write from sock_fd */
close(sock_fd);
```

connection setup: client, using addrinfo

```
int sock_fd;
struct addrinfo *server = /* code on next slide */;

sock_fd = socket(
    server->ai_family,
     // ai_family = AF_INET (IPv4) or AF_INET6 (IPv6) or ...
    server->ai_socktype,
     // ai_socktype = SOCK_STREAM (bytes) or ...
    server->ai_protocol
     // ai_protocol = 0 (default)
);
if (soc
if (con                                                    0) {
    /* 
}
freeaddrinfo(server);
DoClientStuff(sock_fd); /* read and write from sock_fd */
close(sock_fd);
```

addrinfo contains all information needed to setup socket

set by `getaddrinfo` function (next slide)

handles IPv4 and IPv6

handles DNS names, service names

connection setup: client, using addrinfo

```
int sock_fd;
struct addrinfo *server = /* code on next slide */;

sock_fd = socket(
    server->ai_family,
     // ai_family = AF_INET (IPv4) or AF_INET6 (IPv6) or ...
    server->ai_socktype,
     // ai_socktype = SOCK_STREAM (bytes) or ...
    server->ai_prototocol
     // ai_protocol = IPPROTO_TCP or ...
);
if (sock_fd < 0) { /* handle error */ }
if (connect(sock_fd, server->ai_addr, server->ai_addrlen) < 0) {
    /* handle error */
}
freeaddrinfo(server);
DoClientStuff(sock_fd); /* read and write from sock_fd */
close(sock_fd);
```

connection setup: client, using addrinfo

```
int sock_fd;
struct addri
sock_fd = so
    server->
     // ai_family = AF_INET (IPv4) or AF_INET6 (IPv6) or ...
    server->ai_socktype,
     // ai_socktype = SOCK_STREAM (bytes) or ...
    server->ai_prototcol
     // ai_protocol = IPPROTO_TCP or ...
);
if (sock_fd < 0) { /* handle error */ }
if (connect(sock_fd, server->ai_addr, server->ai_addrlen) < 0) {
    /* handle error */
}
freeaddrinfo(server);
DoClientStuff(sock_fd); /* read and write from sock_fd */
close(sock_fd);
```

ai_addr points to struct representing address
type of struct depends whether IPv6 or IPv4

connection setup: client, using addrinfo

```
int sock_fd;
str
soc
      // ai_family = AF_INET (IPv4) or AF_INET6 (IPv6) or ...
    server->ai_socktype,
      // ai_socktype = SOCK_STREAM (bytes) or ...
    server->ai_protocol
      // ai_protocol = IPPROTO_TCP or ...
);
if (sock_fd < 0) { /* handle error */ }
if (connect(sock_fd, server->ai_addr, server->ai_addrlen) < 0) {
    /* handle error */
}
freeaddrinfo(server);
DoClientStuff(sock_fd); /* read and write from sock_fd */
close(sock_fd);
```

since addrinfo contains pointers to dynamically allocated memory, call this function to free everything

connection setup: lookup address

```
/* example hostname, portname = "www.cs.virginia.edu", "443" */
const char *hostname; const char *portname;
...
struct addrinfo *server;
struct addrinfo hints;
int rv;
memset(&hints, 0, sizeof(hints));
hints.ai_family = AF_UNSPEC; /* for IPv4 OR IPv6 */
// hints.ai_family = AF_INET4; /* for IPv4 only */

hints.ai_socktype = SOCK_STREAM; /* byte-oriented --- TCP */
rv = getaddrinfo(hostname, portname, &hints, &server);
if (rv != 0) { /* handle error */ }

/* eventually freeaddrinfo(result) */
```

connection setup: lookup address

```
/* example hostname, portname = "www.cs.virginia.edu", "443" */
const char *hostname; const char *portname;
...
struct addrinfo *server;
struct addrinfo hints;
int rv;
memset(&hints, 0, sizeof(hints));
hints.ai_family = AF_UNSPEC; /* for IPv4 OR IPv6 */
// hints.ai_family = AF_INET; /* for IPv4 only */
hints.ai_socktype = SOCK_STREAM; /* byte-oriented --- TCP */
rv = getaddrinfo(hostname, portname, &hints, &server);
if (rv != 0) { /* handle error */ }

/* eventually freeaddrinfo(result) */
```

NB: pass pointer *to pointer* to addrinfo to fill in

connection setup: lookup address

```
/* example hostname, portname = "www.cs.virginia.edu", "443" */
const c
...
struct
struct
int rv;
memset(&hints, 0, sizeof(hints));
hints.ai_family = AF_UNSPEC; /* for IPv4 OR IPv6 */
// hints.ai_family = AF_INET4; /* for IPv4 only */

hints.ai_socktype = SOCK_STREAM; /* byte-oriented --- TCP */
rv = getaddrinfo(hostname, portname, &hints, &server);
if (rv != 0) { /* handle error */ }

/* eventually freeaddrinfo(result) */
```

AF_UNSPEC: choose between IPv4 and IPv6 for me
AF_INET, AF_INET6: choose IPv4 or IPV6 respectively

connection setup: server, address setup

```
/* example (hostname, portname) = ("127.0.0.1", "443") */
const char *hostname; const char *portname;
...
struct addrinfo *server;
struct addrinfo hints;
int rv;

memset(&hints, 0, sizeof(hints));
hints.ai_family = AF_INET; /* for IPv4 */
/* or: */ hints.ai_family = AF_INET6; /* for IPv6 */
/* or: */ hints.ai_family = AF_UNSPEC; /* I don't care */
hints.ai_flags = AI_PASSIVE;

rv = getaddrinfo(hostname, portname, &hints, &server);
if (rv != 0) { /* handle error */ }
```

connection setup: server, address setup

```
/* example (hostname, portname) = ("127.0.0.1", "443") */
const char *hostname; const char *portname;
...
struct addrinfo *server;
struct addrinfo hints;
int rv;

memset(&hints, 0, sizeof(hints));
hints.ai_family = AF_INET; /* for IPv4 */
/* or: */ hints.ai_family = AF_INET6; /* for IPv6 */
/* or: */ hints.ai_family = AF_UNSPEC; /* I don't care */
hints.ai_flags =

rv = getaddrinfo(                                    ver);
if (rv != 0) { /*
```

hostname could also be NULL
means “use all possible addresses”
only makes sense for servers

connection setup: server, address setup

```
/* example (hostname, portname) = ("127.0.0.1", "443") */
const char *hostname; const char *portname;
...
struct addrinfo *server;
struct addrinfo hints;
int rv;

memset(&hints, 0, sizeof(hints));
hints.ai_family = AF_INET; /* for IPv4 */
/* or: */ hints.ai_family = AF_INET6; /* for IPv6 */
/* or: */ hints.ai_family = AF_UNSPEC; /* I don't care */
hints.ai_flags

rv = getaddrinf                                    r);
if (rv != 0) {
```

portname could also be NULL
means “choose a port number for me”
only makes sense for servers

connection setup: server, address setup

```
/* example (hostname, portname) = ("127.0.0.1", "443") */
const char *hos
...
struct addrinfo *server;
struct addrinfo hints;
int rv;

memset(&hints, 0, sizeof(hints));
hints.ai_family = AF_INET; /* for IPv4 */
/* or: */ hints.ai_family = AF_INET6; /* for IPv6 */
/* or: */ hints.ai_family = AF_UNSPEC; /* I don't care */
hints.ai_flags = AI_PASSIVE;

rv = getaddrinfo(hostname, portname, &hints, &server);
if (rv != 0) { /* handle error */ }
```

AI_PASSIVE: “I’m going to use bind”

connection setup: server, addrinfo

```
struct addrinfo *server;
... getaddrinfo(...) ...

int server_socket_fd = socket(
    server->ai_family,
    server->ai_sockttype,
    server->ai_protocol
);

if (bind(server_socket_fd, ai->ai_addr, ai->ai_addr_len)) < 0) {
    /* handle error */
}
listen(server_socket_fd, MAX_NUM_WAITING);
...
int socket_fd = accept(server_socket_fd, NULL);
```

aside: on server port numbers

Unix convention: must be `root` to use ports 0–1023

root = superuser = 'adminstrator user' = what sudo does

so, for testing: probably ports > 1023