?????????

v1

* ???

github.com/rfyiamcool

1 socket????

2 io ????

3 ??????

4 ????

1

socket

socket

????

????????????socket??????;

?? socket waitqueue ??????????????.

????

??????

??ringbuffer

?? vs ???

????

wait for data !!!

return data

?????

???

????

??/??? IO

????

????

??/?? IO

????

????

?? vs ??

??
??????? socket ??????????????????????????

??
???????????????????????,??????

?? IO ??

??io??

???io??

io??????????IO????io??

?? io ??

??????????, ??????????.
?????????????, ?????????.??
???????????
??

???????????fd

??? io ??

socket ?????, ????????, ?? cpu ??

??????, ?????, ?????? EAGAIN ? EWOULDBLOCK

??????, ???????, ????? EAGAIN ? EWOULDBLOCK

???

??????????

??

???, ????

io ??????

lib
select poll epoll

??????????
????, ??????, ???? block ??
? kernel ??????????recvfrom ?????
kernel?????????????????block?

aio ??

aio_read??????, kernel???????????????????????, ?????????? ;

?????? block ???????, ??????? read ?? ;

2

IO ????

select

?????????, 32bit ? 1024 , 64bit ? 2048

fd_set ????, ?????????
??????????? fd_set

O(n) ???????????????

poll

??? select fd_size ??

?????? fd_set ?????fd???????

???? ???? ????

epoll

????????

????????, ??????, ?????? ready list ??? epitem ????????????

??????/??/????

epoll

epfd = epoll_create(intsize);

epoll_ctl(int epfd, int op, int fd, struct epoll_event *event) EPOLL_CTL_ADD EPOLL_CTL_MOD EPOLL_CTL_DEL

epoll_wait(int epfd, struct epoll_event * events, intmaxevents, int timeout) ;

epoll event

?? Event

EPOLLIN ???

EPOLLOUT ???

EPOLLET ??????

EPOLLLT ??????

EPOLLRDHUP, EPOLLHUP ????

EPOLLERR ????

epoll desgin

2.3 ????? wq ?????????fd??

2.2 ?? callback ???????
copy ? rdlist

epoll

???? (level trigger)

??????
???? (???????) ,?? (??????) ?????????

???????, ???? epollout ??, ?????? !

???? (edge trigger)

????????????, ????????? .
??????, ????????????? lt ??

???????http??, ?????????, ??????????, ???fd?????? !

epoll

???? lt

- java nio
- redis
- muduo
- skynet
- tornado
- golang gev
- libevent (default)

???? et

- nginx
- envoy
- mysql
- netty (default)
- golang net
- golang gnet

epoll

epoll ????? ?

epoll ?? mmap ???? ? epoll ???? ?

epoll ???? fd ?? ?

epoll ??????

??????? ?

sockfd

- eventfd
- timerfd
- signalfd
- inotifyfd
- pipefd

AIO (async io)

kernel native aio

????
nginx, mysql ????
???? direct io ??io??? socket !!!

glibc aio

??????? user ???io

io_uring

io_uring (??????????)
kernel 5.1 ??, kernel 5.6 ????? epoll

??????
??socket fd mmap ??
?? buffered io

3

??????

single thread

new thread

??
???????????

thread pool

??????? + ????????
????

eventloop ???????????
?????conn????????

????????, ?????
? poller ??????

multiple worker

???? worker ??/????? .

?? master ?????? worker ?? .

?????????? event-loop !!!

nginx, haproxy, apache ????? .

reactor

java netty
c++ muduo
golang gnet
golang gev

?.

reactor

OnConnection()
OnMessage()
OnClose()

connection
connection
connection
connection
connection

listen fd

dispatch algo
rr lc p2c

??ing

http request upstream backend
query mysql, redis, http api block
queue
mutex / semaphore disk io

?????? ?/? ???

proactor

??? aio_read ??? I/O? ??, ?????,?????, ???????????????????, ???????????????? !!!

??????,
?????????????

Reactor, ???????????, ???????????

Proactor, ???????, ????????????

?? AIO ?? !!!

proactor

linux aio ???? !!!

????linux???? reactor ??

???? socket aio, ??????? aio

?????? direct io

windows/darwin ??????? aio, ?????? proactor ??

so_reuseaddr

?????????close??, ?????????? time-wait ??

socket ??? time-wait ??????bind??????.?? : ?Address already in use?

so_reuseport

????

fork

fd table

listen fd

fd table

listen fd

??? accept ????? kernel 2.6 ?? !

?? eventloop ?????? fd ??epoll??

fork

????

epoll listen fd ?

?accept_mutex? , ?? nginx ???
reuseport ??????, ??worker?????? listen fd epoll exclusive on Linux 4.5 +

????

?????????????? worker ??

reactor ????

?????

EINTR, ?????????
EAGAIN, ??????
EWOULDBLOCK, ?? (????????) ECONNREST, tcp rst????
EPIPE, ???? ETIMEDOUT, ???? EINPROGRESS, ????
?

N

> 0, ?????N??????
= 0, ????????
-1, ?? errno ??

??????

socket ?????, ?? connect ??

? ret ??0??????, ? -1 ? errno = EINPROGRESS ?????

?? fd ??select??epoll??? EPOLLOUT ??

??????,
???getsockopt???status?????0?????connect??????0??????????????ip????

4 ????

tcp/udp ??????

????? ?

?????????? ? udp ??????? ?

????

???????.

???????.

??????

mtu/mss (1500/1460)
nagle rto / ofo

??, ????, ?????, ????????

network delay

???? socket.send()

??, ????, ??????????

??????????
. . .

????

tlv

golang pack/unpack

????

udp ????????? !!!

udp ??????, ???????

???????? udp ???????

????? 10 ???, ???????10?

?????????, ?????

??? udp ? tlv ??, ?????? !!!

????

?????????

- ??????? socket ??
- golang ??????????
- tcp keepalive ????????? ?
- ?????????

??????

sysctl -a

net.ipv4.tcp_keepalive_time = 7200

??7200?????????
net.ipv4.tcp_keepalive_intvl = 75
???? 75s net.ipv4.tcp_keepalive_probes = 9

???9?
???? 2 ?? 11 ? 15 ? ???????? !

nima, ????? !!!

??????

???? sysctl ? keepalive ???? ;

??????? socket opt ???? ;

????????? ;

??????? !!!

golang ?????

golang ????? socket ?? 15s ???

?? socket setsockopt ??

?????

???????????

tcp keepalive ???????????????????????

tcp keepalive ??????????????????, ?????????TCP?????????????????????????? ;

tcp keepalive ?????????????? ;

tcp time-wait ?????

????? timewait ?

timewait ????? ?????

timewait ? ???? tiemwait ?

timewait

timewait ?????? ?
??????????? timewait !!!

FIN_WAIT_2

CLOSE

time-wait

??????

???????time-wait

!!!

time-wait ??

wan ip

5:01

5:05

????

?????

?????

????

syn queue

accept queue

????

backlog?net.ipv4.tcp_max_syn_backlog?
net.core.somaxconn (???????)

Redis = 511
net.core.somaxconn (128)
accept????????
????????????????????

????

connection reset by peer &
connection refused

?????syn, ?????????
????????syn, ??? tcp_syn_retries = 5

??????? ?

?????ack, ??????????syn/ack

??? tcp_synack_retries = 5

equal 1

retrun rst

????????? udp !!!

???
??????

????

????????????IP ?

????????

?????, ????? RTT ????
??????????????

IP ???????????? IP !!!
(???????????)

???????? ?

?????? ?

????????????

????

?????ack, cwnd = cwnd +1

???RTT, cwnd = cwnd x 2

????

cwnd > ssthresh???????????RTT, cwnd = cwnd + 1

????

????

threhold = cwnd/2 , cwnd = 1

???????????

cwnd = cwnd /2, threhold = cwnd

??????

????

cwnd = threhold + 3 MSS

kcp

??udp???????

github.com/skywind3000/kcp

protocol over quic

?? UDP ??

????? (??, ????)

?? 0-RTT
FEC ?????????

???????? HOL

connection migration

connection id

123456

connection id

123456

protocol over quic

mqtt over quic
http over quic grpc
over quic rpcx over
quic

?

????

?? IO vs ?? IP vs UDS (Unix Domain Socket)
client/server ??????? ? client/server ? crash ????? ? socket buffer
????

local ip vs uds kernel path

unix socket domain

local ip

???? ?

uds ?????

local ip ????

uds ???? localhost ???

?????? ?

??????? (ulimit)
???? (buffer/cache) other sysctl.conf

???

?????????????tcp 4 ??

client?ip?????2?32??, port?2?16??
???????????2?48?????

?????????????tcp 4 ??

ip_local_port_range (default 32768 - 60999) peer to peer

192.168.0.99

client

????

bind local interface

192.168.0.99

client

????

dial more server ports

192.168.0.99

client

dial more servers

when process crash ?

- ??????, ?????????? ?
  - ?????
    - ????????? fin
  - ?????
    - ?? rst
  - linger
    - ?? rst

?????? !!!

- FIN
  - init 0 (??)
  - kill -15 pid
  - kill -9 pid
  - OOM without recv-q

when server crash ?

??? curl/telnet ??? ?

????????? ?

????????? rst

?????????? ?

?????, ????

when router crash ?

???????????? ?

tcp???????????? !!!

buffer

????? socket ?? SO_SNDBUF, SO_RCVBUF ???

linux ???????????????

???? BDP ?????? BDP=rtt*(??/8) ?????

grace upgrade

????
fork exec

system = fork + exec + waitpid

????? close-on-exec flag, ?????????????? !!!

grace upgrade

?? listenfd
nginx & golang by dup envoy by
uds

if don?t recv quit signal for a long time,
send kill !

golang ?????

https://github.com/openimsdk/open-im-server
https://github.com/tidwall/evio https://github.com/panjf2000/gnet
https://github.com/lesismal/nbio https://github.com/Allenxuxu/gev
https://github.com/shaovie/goev

Q & A

- xiaorui.cc
- github.com/rfyiamcool