Linux x86_64系统调用的实现过程

Author: Lan Tianyu, Date: 2019-03-14 19:20:00 +0800, Last modified: 2019-03-14 19:20:00 +0800

Origin: https://blog.lantianyu.com/2019/03-14-1920-linux-x86-64-system-call.html

Copyright: 本文采用 CC BY-NC-SA 3.0 进行授权

demo

```
[root@localhost ~]# cat hello.s
.section .rodata
.LC0:
        .string "hello\n"
.section .text
.global _start
_start:
        movq $1, %rax
        movq $1, %rdi
        leaq .LC0(%rip), %rsi
        movq $6, %rdx
        syscall
        movq $60, %rax
        movq $0, %rdi
        syscall
[root@localhost ~]# gcc -c hello.s
[root@localhost ~]# ld -o hello hello.o
[root@localhost ~]# ./hello
hello
[root@localhost ~]# strace ./hello
execve("./hello", [], [/* 27 vars */]) = 0
write(1, "hello\n", 6)                  = 6
exit_syscall(0)                         = ?
+++ exited with 0 +++
[root@localhost ~]# gdb -q ./hello
Reading symbols from ./hello...(no debugging symbols found)...done.
(gdb) b _start
Breakpoint 1 at 0x401000
(gdb) r
Starting program: /root/hello
Breakpoint 1, 0x0000000000401000 in _start ()
(gdb) disassemble _start
Dump of assembler code for function _start:
=> 0x0000000000401000:  mov    $0x1,%rax
   0x0000000000401007:  mov    $0x1,%rdi
   0x000000000040100e:  lea    0x0(%rip),%rsi        # 0x401015
   0x0000000000401015:  mov    $0x6,%rdx
   0x000000000040101c:  syscall
   0x000000000040101e:  mov    $0x3c,%rax
   0x0000000000401025:  mov    $0x0,%rdi
   0x000000000040102c:  syscall
End of assembler dump.
(gdb) si
0x0000000000401007 in _start ()
(gdb) si
0x000000000040100e in _start ()
(gdb) si
0x0000000000401015 in _start ()
(gdb) si
0x000000000040101c in _start ()
(gdb) si
hello
0x000000000040101e in _start ()
(gdb) si
0x0000000000401025 in _start ()
(gdb) si
0x000000000040102c in _start ()
(gdb) si
[Inferior is 1 (process 3000) exited with code 00]
(gdb) q
[root@localhost ~]#
```

上面的一个demo，使用了strace命令来跟踪程序的系统调用。strace命令的输出表明，程序调用了两个系统调用，分别是write和exit_syscall。

gdb的disassemble命令，可以反汇编出程序的汇编语言。

gdb的si命令，可以单步执行程序的汇编指令。

二、分析一个系统调用的源码

Linux内核版本：4.19.0-0.bpo.2-amd64

```
[root@localhost ~]# cat /usr/src/linux-4.19.0-0.bpo.2-amd64/arch/x86/entry/syscalls/syscall_64.tbl
# 64-bit system call numbers and entry vectors
#
# The format is
# <number> <abi> <name> <entry_point>
#
# The abi is "common", "64" or "x32" for this file.
#
0       common  read            sys_read
1       common  write           sys_write
2       common  open            sys_open
3       common  close           sys_close
4       common  stat            sys_newstat
5       common  fstat           sys_newfstat
6       common  lstat           sys_newlstat
7       common  poll            sys_poll
8       common  lseek           sys_lseek
9       common  mmap            sys_mmap
10      common  mprotect        sys_mprotect
11      common  munmap          sys_munmap
12      common  brk             sys_brk
13      64      rt_sigaction    sys_rt_sigaction
14      common  rt_sigprocmask  sys_rt_sigprocmask
15      64      rt_sigreturn    stub_rt_sigreturn
16      64      ioctl           sys_ioctl
17      common  pread           sys_pread64
18      common  pwrite          sys_pwrite64
19      64      readv           sys_readv
20      common  writev          sys_writev
21      common  access          sys_access
22      common  pipe            sys_pipe
23      common  select          sys_select
24      common  sched_yield     sys_sched_yield
25      common  mremap          sys_mremap
26      common  msync           sys_msync
27      common  mincore         sys_mincore
28      common  madvise         sys_madvise
29      common  shmget          sys_shmget
30      common  shmat           sys_shmat
31      common  shmctl          sys_shmctl
32      common  dup             sys_dup
33      common  dup2            sys_dup2
34      common  pause           sys_pause
35      common  nanosleep       sys_nanosleep
36      common  getitimer       sys_getitimer
37      common  alarm           sys_alarm
38      common  setitimer       sys_setitimer
39      common  getpid          sys_getpid
40      common  sendfile        sys_sendfile64
41      common  socket          sys_socket
42      common  connect         sys_connect
43      common  accept          sys_accept
44      common  sendto          sys_sendto
45      common  recvfrom        sys_recvfrom
46      64      sendmsg         sys_sendmsg
47      64      recvmsg         sys_recvmsg
48      common  shutdown        sys_shutdown
49      common  bind            sys_bind
50      common  listen          sys_listen
51      common  getsockname     sys_getsockname
52      common  getpeername     sys_getpeername
53      common  socketpair      sys_socketpair
54      64      setsockopt      sys_setsockopt
55      64      getsockopt      sys_getsockopt
56      common  clone           stub_clone
57      common  fork            stub_fork
58      common  vfork           stub_vfork
59      64      execve          stub_execve
60      common  exit            sys_exit
61      common  wait4           sys_wait4
62      common  kill            sys_kill
63      common  uname           sys_newuname
64      common  semget          sys_semget
65      common  semop           sys_semop
66      common  semctl          sys_semctl
67      common  shmdt           sys_shmdt
68      common  msgget          sys_msgget
69      common  msgsnd          sys_msgsnd
70      common  msgrcv          sys_msgrcv
71      common  msgctl          sys_msgctl
72      common  fcntl           sys_fcntl
73      common  flock           sys_flock
74      common  fsync           sys_fsync
75      common  fdatasync       sys_fdatasync
76      common  truncate        sys_truncate
77      common  ftruncate       sys_ftruncate
78      common  getdents        sys_getdents
79      common  getcwd          sys_getcwd
80      common  chdir           sys_chdir
81      common  fchdir          sys_fchdir
82      common  rename          sys_rename
83      common  mkdir           sys_mkdir
84      common  rmdir           sys_rmdir
85      common  creat           sys_creat
86      common  link            sys_link
87      common  unlink          sys_unlink
88      common  symlink         sys_symlink
89      common  readlink        sys_readlink
90      common  chmod           sys_chmod
91      common  fchmod          sys_fchmod
92      common  chown           sys_chown
93      common  fchown          sys_fchown
94      common  lchown          sys_lchown
95      common  umask           sys_umask
96      common  gettimeofday    sys_gettimeofday
97      common  getrlimit       sys_getrlimit
98      common  getrusage       sys_getrusage
99      common  sysinfo         sys_sysinfo
100     common  times           sys_times
101     64      ptrace          sys_ptrace
102     common  getuid          sys_getuid
103     common  syslog          sys_syslog
104     common  getgid          sys_getgid
105     common  setuid          sys_setuid
106     common  setgid          sys_setgid
107     common  geteuid         sys_geteuid
108     common  getegid         sys_getegid
109     common  setpgid         sys_setpgid
110     common  getppid         sys_getppid
111     common  getpgrp         sys_getpgrp
112     common  setsid          sys_setsid
113     common  setreuid        sys_setreuid
114     common  setregid        sys_setregid
115     common  getgroups       sys_getgroups
116     common  setgroups       sys_setgroups
117     common  setresuid       sys_setresuid
118     common  getresuid       sys_getresuid
119     common  setresgid       sys_setresgid
120     common  getresgid       sys_getresgid
121     common  getpgid         sys_getpgid
122     common  setfsuid        sys_setfsuid
123     common  setfsgid        sys_setfsgid
124     common  getsid          sys_getsid
125     common  capget          sys_capget
126     common  capset          sys_capset
127     64      rt_sigpending   sys_rt_sigpending
128     64      rt_sigtimedwait sys_rt_sigtimedwait
129     64      rt_sigqueueinfo sys_rt_sigqueueinfo
130     common  rt_sigsuspend   sys_rt_sigsuspend
131     64      sigaltstack     sys_sigaltstack
132     common  utime           sys_utime
133     common  mknod           sys_mknod
134     64      uselib          
135     common  personality     sys_personality
136     common  ustat           sys_ustat
137     common  statfs          sys_statfs
138     common  fstatfs         sys_fstatfs
139     64      sysfs           sys_sysfs
140     common  getpriority     sys_getpriority
141     common  setpriority     sys_setpriority
142     common  sched_setparam  sys_sched_setparam
143     common  sched_getparam  sys_sched_getparam
144     common  sched_setscheduler sys_sched_setscheduler
145     common  sched_getscheduler sys_sched_getscheduler
146     common  sched_get_priority_max sys_sched_get_priority_max
147     common  sched_get_priority_min sys_sched_get_priority_min
148     common  sched_rr_get_interval sys_sched_rr_get_interval
149     common  mlock           sys_mlock
150     common  munlock         sys_munlock
151     common  mlockall        sys_mlockall
152     common  munlockall      sys_munlockall
153     common  vhangup         sys_vhangup
154     common  modify_ldt      sys_modify_ldt
155     common  pivot_root      sys_pivot_root
156     64      _sysctl         sys_ni_syscall
157     common  prctl           sys_prctl
158     common  arch_prctl      sys_arch_prctl
159     common  adjtimex        sys_adjtimex
160     common  setrlimit       sys_setrlimit
161     common  chroot          sys_chroot
162     common  sync            sys_sync
163     common  acct            sys_acct
164     common  settimeofday    sys_settimeofday
165     common  mount           sys_mount
166     common  umount2         sys_umount
167     common  swapon          sys_swapon
168     common  swapoff         sys_swapoff
169     common  reboot          sys_reboot
170     common  sethostname     sys_sethostname
171     common  setdomainname   sys_setdomainname
172     common  iopl            stub_iopl
173     64      ioperm          sys_ioperm
174     64      create_module   
175     common  init_module     sys_init_module
176     common  delete_module   sys_delete_module
177     64      get_kernel_syms 
178     64      query_module    
179     common  quotactl        sys_quotactl
180     64      nfsservctl      
181     common  getpmsg         
182     common  putpmsg         
183     common  afs_syscall     
184     common  tuxcall         
185     64      security        
186     common  gettid          sys_gettid
187     common  readahead       sys_readahead
188     common  setxattr        sys_setxattr
189     common  lsetxattr       sys_lsetxattr
190     common  fsetxattr       sys_fsetxattr
191     common  getxattr        sys_getxattr
192     common  lgetxattr       sys_lgetxattr
193     common  fgetxattr       sys_fgetxattr
194     common  listxattr       sys_listxattr
195     common  llistxattr      sys_llistxattr
196     common  flistxattr      sys_flistxattr
197     common  removexattr     sys_removexattr
198     common  lremovexattr    sys_lremovexattr
199     common  fremovexattr    sys_fremovexattr
200     common  tkill           sys_tkill
201     common  time            sys_time
202     common  futex           sys_futex
203     common  sched_setaffinity sys_sched_setaffinity
204     common  sched_getaffinity sys_sched_getaffinity
205     64      set_thread_area 
206     64      io_setup        sys_io_setup
207     common  io_destroy      sys_io_destroy
208     common  io_getevents    sys_io_getevents
209     64      io_submit       sys_io_submit
210     common  io_cancel       sys_io_cancel
211     64      get_thread_area 
212     common  lookup_dcookie  sys_lookup_dcookie
213     common  epoll_create    sys_epoll_create
214     64      epoll_ctl_old   
215     64      epoll_wait_old  
216     common  remap_file_pages sys_remap_file_pages
217     common  getdents64      sys_getdents64
218     common  set_tid_address sys_set_tid_address
219     common  restart_syscall sys_restart_syscall
220     64      semtimedop      sys_semtimedop
221     common  fadvise64       sys_fadvise64
222     64      timer_create    sys_timer_create
223     common  timer_settime   sys_timer_settime
224     common  timer_gettime   sys_timer_gettime
225     common  timer_getoverrun sys_timer_getoverrun
226     common  timer_delete    sys_timer_delete
227     common  clock_settime   sys_clock_settime
228     common  clock_gettime   sys_clock_gettime
229     common  clock_getres    sys_clock_getres
230     common  clock_nanosleep sys_clock_nanosleep
231     common  exit_group      sys_exit_group
232     common  epoll_wait      sys_epoll_wait
233     common  epoll_ctl       sys_epoll_ctl
234     common  tgkill          sys_tgkill
235     common  utimes          sys_utimes
236     64      vserver         
237     common  mbind           sys_mbind
238     common  set_mempolicy   sys_set_mempolicy
239     common  get_mempolicy   sys_get_mempolicy
240     common  mq_open         sys_mq_open
241     common  mq_unlink       sys_mq_unlink
242     common  mq_timedsend    sys_mq_timedsend
243     common  mq_timedreceive sys_mq_timedreceive
244     64      mq_notify       sys_mq_notify
245     common  mq_getsetattr   sys_mq_getsetattr
246     64      kexec_load      sys_kexec_load
247     32      waitid          sys_waitid
248     common  add_key         sys_add_key
249     common  request_key     sys_request_key
250     common  keyctl          sys_keyctl
251     common  ioprio_set      sys_ioprio_set
252     common  ioprio_get      sys_ioprio_get
253     common  inotify_init    sys_inotify_init
254     common  inotify_add_watch sys_inotify_add_watch
255     common  inotify_rm_watch sys_inotify_rm_watch
256     common  migrate_pages   sys_migrate_pages
257     common  openat          sys_openat
258     common  mkdirat         sys_mkdirat
259     common  mknodat         sys_mknodat
260     common  fchownat        sys_fchownat
261     common  futimesat       sys_futimesat
262     common  newfstatat      sys_newfstatat
263     common  unlinkat        sys_unlinkat
264     common  renameat        sys_renameat
265     common  linkat          sys_linkat
266     common  symlinkat       sys_symlinkat
267     common  readlinkat      sys_readlinkat
268     common  fchmodat        sys_fchmodat
269     common  faccessat       sys_faccessat
270     common  pselect6        sys_pselect6
271     common  ppoll           sys_ppoll
272     common  unshare         sys_unshare
273     64      set_robust_list sys_set_robust_list
274     64      get_robust_list sys_get_robust_list
275     common  splice          sys_splice
276     common  tee             sys_tee
277     common  sync_file_range sys_sync_file_range
278     64      vmsplice        sys_vmsplice
279     64      move_pages      sys_move_pages
280     common  utimensat       sys_utimensat
281     common  epoll_pwait     sys_epoll_pwait
282     common  signalfd        sys_signalfd
283     common  timerfd_create  sys_timerfd_create
284     common  eventfd         sys_eventfd
285     common  fallocate       sys_fallocate
286     common  timerfd_settime sys_timerfd_settime
287     common  timerfd_gettime sys_timerfd_gettime
288     common  accept4         sys_accept4
289     common  signalfd4       sys_signalfd4
290     common  eventfd2        sys_eventfd2
291     common  epoll_create1   sys_epoll_create1
292     common  dup3            sys_dup3
293     common  pipe2           sys_pipe2
294     common  inotify_init1   sys_inotify_init1
295     64      preadv          sys_preadv
296     64      pwritev         sys_pwritev
297     64      rt_tgsigqueueinfo sys_rt_tgsigqueueinfo
298     common  perf_event_open sys_perf_event_open
299     64      recvmmsg        sys_recvmmsg
300     common  fanotify_init   sys_fanotify_init
301     common  fanotify_mark   sys_fanotify_mark
302     common  prlimit64       sys_prlimit64
303     common  name_to_handle_at sys_name_to_handle_at
304     common  open_by_handle_at sys_open_by_handle_at
305     common  clock_adjtime   sys_clock_adjtime
306     common  syncfs          sys_syncfs
307     64      sendmmsg        sys_sendmmsg
308     common  setns           sys_setns
309     common  getcpu          sys_getcpu
310     64      process_vm_readv sys_process_vm_readv
311     64      process_vm_writev sys_process_vm_writev
312     common  kcmp            sys_kcmp
313     common  finit_module    sys_finit_module
314     common  sched_setattr   sys_sched_setattr
315     common  sched_getattr   sys_sched_getattr
316     common  renameat2       sys_renameat2
317     common  seccomp         sys_seccomp
318     common  getrandom       sys_getrandom
319     common  memfd_create    sys_memfd_create
320     common  kexec_file_load sys_kexec_file_load
321     common  bpf             sys_bpf
322     64      execveat        stub_execveat
323     common  userfaultfd     sys_userfaultfd
324     common  membarrier      sys_membarrier
325     common  mlock2          sys_mlock2
326     common  copy_file_range sys_copy_file_range
327     64      preadv2         sys_preadv2
328     64      pwritev2        sys_pwritev2
329     common  pkey_mprotect   sys_pkey_mprotect
330     common  pkey_alloc      sys_pkey_alloc
331     common  pkey_free       sys_pkey_free
332     common  statx           sys_statx
333     common  io_pgetevents   sys_io_pgetevents
334     common  rseq            sys_rseq
[root@localhost ~]#
```

上面的syscall_64.tbl文件，是Linux内核源代码中，定义系统调用的文件。

syscall_64.tbl文件的第一列，是系统调用的编号。

第二列，是调用的方式。

第三列，是系统调用的名称。

第四列，是系统调用的入口函数。

```
[root@localhost ~]# cat /usr/src/linux-4.19.0-0.bpo.2-amd64/arch/x86/entry/syscalls/syscallhdr.sh
[root@localhost ~]# /usr/src/linux-4.19.0-0.bpo.2-amd64/arch/x86/entry/syscalls/syscallhdr.sh /usr/src/linux-4.19.0-0.bpo.2-amd64/arch/x86/entry/syscalls/syscall_64.tbl __NR_syscallhdr.tmp
#!/usr/bin/awk -f
# SPDX-License-Rev: 1.0
# Copyright (C) 2017-2018 Jason A. Donenfeld <Jason@zx2c4.com>. All Rights Reserved.
# This script generates a header file containing system call numbers.
# It parses the system call table, and optionally the system call prototypes.
# Usage: syscallhdr.sh --tbl <table file> [--prefix <prefix>] [--emit-nr] [--emit-proto] [--emit-guard] [--emit-table] [--emit-initializer] [--emit-unistd] [--emit-unistd-proto] [--emit-unistd-table] [--emit-unistd-initializer] [--emit-unistd-guard] [--emit-unistd-nr] [--emit-unistd-proto-nr] [--emit-unistd-table-nr] [--emit-unistd-initializer-nr] [--emit-unistd-guard-nr] [--emit-unistd-proto-table] [--emit-unistd-proto-initializer] [--emit-unistd-proto-guard] [--emit-unistd-table-initializer] [--emit-unistd-table-guard] [--emit-unistd-initializer-guard] [--emit-unistd-proto-table-initializer] [--emit-unistd-proto-table-guard] [--emit-unistd-proto-initializer-guard] [--emit-unistd-table-initializer-guard] [--emit-unistd-proto-table-initializer-guard] [--emit-unistd-proto-table-initializer-guard-nr] [--emit-unistd-proto-table-initializer-guard-nr-table] [--emit-unistd-proto-table-initializer-guard-nr-table-initializer] [--emit-unistd-proto-table-initializer-guard-nr-table-initializer-guard] [--emit-unistd-proto-table-initializer-guard-nr-table-initializer-guard-nr] [--emit-unistd-proto-table-initializer-guard-nr-table-initializer-guard-nr-table] [--emit-unistd-proto-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer] [--emit-unistd-proto-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard] [--emit-unistd-proto-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr] [--emit-unistd-proto-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table] [--emit-unistd-proto-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer] [--emit-unistd-proto-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard] [--emit-unistd-proto-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr] [--emit-unistd-proto-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table] [--emit-unistd-proto-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer] [--emit-unistd-proto-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard] [--emit-unistd-proto-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr] [--emit-unistd-proto-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table] [--emit-unistd-proto-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer] [--emit-unistd-proto-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard] [--emit-unistd-proto-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr] [--emit-unistd-proto-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table] [--emit-unistd-proto-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer] [--emit-unistd-proto-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard] [--emit-unistd-proto-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr] [--emit-unistd-proto-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table] [--emit-unistd-proto-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer] [--emit-unistd-proto-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard] [--emit-unistd-proto-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr] [--emit-unistd-proto-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table] [--emit-unistd-proto-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer-guard-nr-table-initializer
```