
Subject: Re: Hang with fair cgroup scheduler (reproducer is attached.)

Posted by [Ingo Molnar](#) on Fri, 14 Dec 2007 09:49:09 GMT

[View Forum Message](#) <> [Reply to Message](#)

(Cc:-ed other folks as well)

* KAMEZAWA Hiroyuki <kamezawa.hiroyu@jp.fujitsu.com> wrote:

> Tested again, and got NULL access and panic.

>

> This is my guess from stack dump. (raw stack dump is attached below.)

> ==

>

> static struct task_struct *pick_next_task_fair(struct rq *rq)

> {

> struct cfs_rq *cfs_rq = &rq->cfs;

> struct sched_entity *se;

>

> if (unlikely(!cfs_rq->nr_running))

> return NULL;

>

> do {

> se = pick_next_entity(cfs_rq); <-- se was NULL.

> cfs_rq = group_cfs_rq(se); <-- se->my_q causes SEGV

> } while (cfs_rq);

>

> return task_of(se);

> }

> ===

> Seems first_fair() was NULL in

> ==

> static struct sched_entity *pick_next_entity(struct cfs_rq *cfs_rq)

> {

> struct sched_entity *se = NULL;

>

> if (first_fair(cfs_rq)) { <-----(*)

> se = __pick_next_entity(cfs_rq);

> set_next_entity(cfs_rq, se);

> }

>

> return se;

> }

> ==

> from register information.

>

> Thanks,

> -Kame

>

```

>
> Stack dump is here.
> ==
> Pid: 8197, CPU 6, comm:                reg
> psr : 00001210085a2010 ifs : 8000000000000206 ip : [<a000000100067c01>]  Not tainted
> ip is at pick_next_task_fair+0x81/0xe0
> unat: 0000000000000000 pfs : 0000000000000206 rsc : 0000000000000003
> rnat: 0000000000000000 bsp: 0000000000000000 pr : 0000000000556959
> ldrs: 0000000000000000 ccv : 0000000000000000 fpsr: 0009804c0270033f
> csd : 0000000000000000 ssd : 0000000000000000
> b0 : a000000100067c00 b6 : a000000100076a60 b7 : a00000010000ee50
> NaT consumption 2216203124768 [1]^M
> Modules linked in: sunrpc binfmt_misc dm_mirror dm_mod fan sg thermal e1000 processor
button conta
> iner e100 eeepro100 mii lpfc mptspi mptscsih mptbase ehci_hcd ohci_hcd uhci_hcd^M
> ^M
> Pid: 8197, CPU 6, comm:                reg^M
> psr : 00001210085a2010 ifs : 8000000000000206 ip : [<a000000100067c01>]  Not tainted^M
> ip is at pick_next_task_fair+0x81/0xe0^M
> unat: 0000000000000000 pfs : 0000000000000206 rsc : 0000000000000003^M
> rnat: 0000000000000000 bsp: 0000000000000000 pr : 0000000000556959^M
> ldrs: 0000000000000000 ccv : 0000000000000000 fpsr: 0009804c0270033f^M
> csd : 0000000000000000 ssd : 0000000000000000^M
> b0 : a000000100067c00 b6 : a000000100076a60 b7 : a00000010000ee50^M
> f6 : 00000000000000000000000000000000 f7 : 000000000000000000000000^M
> f8 : 1003e00000000a0000007 f9 : 1003e00000059499dd2c3^M
> f10 : 1003ece02a62ae350c355 f11 : 1003e0000000000000000037^M
> r1 : a000000100d87a60 r2 : 000000df13538d0b r3 : 000000000000000060^M
> r8 : 0000000000000000 r9 : e00001a004034b30 r10 : 0000000000000000^M
> r11 : e00001a004034aa8 r12 : e00001a10397fe10 r13 : e00001a103970000^M
> r14 : 00000000d594bde3 r15 : e00001a004034ab0 r16 : e00001a004034ab8^M
> r17 : e00001a004034ac8 r18 : e00001a004038320 r19 : e00001a10426ff20^M
> r20 : 0000000000000000 r21 : 0000000000000000 r22 : 0000000000000001^M
> r23 : e00001a004034a91 r24 : e00001a004034a90 r25 : e00001a10426ff10^M
> r26 : 00000000000000002 r27 : e00001a0040382f0 r28 : e00001a004038288^M
> r29 : a0000001008a5468 r30 : a000000100076a60 r31 : a000000100b726e0^M
> ^M
> Call Trace:^M
> [<a000000100013bc0>] show_stack+0x40/0xa0^M
>                sp=e00001a10397f860 bsp=e00001a103970f18^M
> [<a000000100014840>] show_regs+0x840/0x880^M
>                sp=e00001a10397fa30 bsp=e00001a103970ec0^M
> [<a000000100036fa0>] die+0x1a0/0x2a0^M
>                sp=e00001a10397fa30 bsp=e00001a103970e78^M
> [<a0000001000370f0>] die_if_kernel+0x50/0x80^M
>                sp=e00001a10397fa30 bsp=e00001a103970e48^M
> [<a000000100038260>] ia64_fault+0x1140/0x1260^M
>                sp=e00001a10397fa30 bsp=e00001a103970de8^M

```

```

> [<a000000010000ae20>] ia64_leave_kernel+0x0/0x270^M
> sp=e00001a10397fc40 bsp=e00001a103970de8^M
> [<a0000000100067c00>] pick_next_task_fair+0x80/0xe0^M
> sp=e00001a10397fe10 bsp=e00001a103970db8^M
> [<a00000001006f6a60>] schedule+0x8e0/0x1280^M
> sp=e00001a10397fe10 bsp=e00001a103970d08^M
> [<a0000000100074e20>] sys_sched_yield+0xe0/0x100^M
> sp=e00001a10397fe30 bsp=e00001a103970ca8^M
> [<a000000010000aca0>] ia64_ret_from_syscall+0x0/0x20^M
> sp=e00001a10397fe30 bsp=e00001a103970ca8^M
> [<a0000000000010720>] __kernel_syscall_via_break+0x0/0x20^M
> sp=e00001a103980000 bsp=e00001a103970ca8^M
>
> Disassemble.
> ==
> a0000000100067b80 <pick_next_task_fair>:
> a0000000100067b80: 18 10 19 08 80 05 [MMB] alloc r34=ar.pfs,6,4,0
> a0000000100067b86: 20 80 83 00 42 00 adds r2=112,r32
> a0000000100067b8c: 00 00 00 20 nop.b 0x0
> a0000000100067b90: 09 20 81 41 00 21 [MMI] adds r36=96,r32
> a0000000100067b96: 00 00 00 02 00 20 nop.m 0x0
> a0000000100067b9c: 04 00 c4 00 mov r33=b0;;
> a0000000100067ba0: 0b 70 00 04 18 10 [MMI] ld8 r14=[r2];;
> a0000000100067ba6: 70 00 38 0c 72 00 cmp.eq p7,p6=0,r14
> a0000000100067bac: 00 00 04 00 nop.i 0x0;;
> a0000000100067bb0: 10 00 00 00 01 c0 [MIB] nop.m 0x0
> a0000000100067bb6: 81 00 00 00 c2 03 (p07) mov r8=r0
> a0000000100067bbc: 80 00 00 41 (p07) br.cond.spnt.few a0000000100067c30
<pick_next_task_fair+0xb
> 0>
> a0000000100067bc0: 09 48 c0 48 00 21 [MMI] adds r9=48,r36
> a0000000100067bc6: 00 00 00 02 00 00 nop.m 0x0
> a0000000100067bcc: 04 00 00 84 mov r32=r0;;
> a0000000100067bd0: 09 00 00 00 01 00 [MMI] nop.m 0x0
> a0000000100067bd6: 80 00 24 30 20 00 ld8 r8=[r9]
> a0000000100067bdc: 00 00 04 00 nop.i 0x0;;
> a0000000100067be0: 03 00 00 00 01 00 [MII] nop.m 0x0
> a0000000100067be6: b0 00 20 14 72 05 cmp.eq p11,p10=0,r8;;
> a0000000100067bec: 04 47 fc 8c (p10) adds r32=-16,r8;;
> a0000000100067bf0: 51 29 01 40 00 21 [MIB] (p10) mov r37=r32
> a0000000100067bf6: 00 00 00 02 00 05 nop.i 0x0
> a0000000100067bfc: 58 fe ff 5a (p10) br.call.dptk.many b0=a0000000100067a40
<set_next_entity>;
> a0000000100067c00: 0b 18 80 41 00 21 [MMI] adds r3=96,r32;;
> a0000000100067c06: 40 02 0c 30 20 00 ld8 r36=[r3] <-----panic.
> a0000000100067c0c: 00 00 04 00 nop.i 0x0;;
> a0000000100067c10: 10 00 00 00 01 00 [MIB] nop.m 0x0
> a0000000100067c16: 90 00 90 10 72 04 cmp.eq p9,p8=0,r36

```

> a000000100067c1c: b0 ff ff 4a (p08) br.cond.dptk.few a000000100067bc0
<pick_next_task_fair+0x4

Containers mailing list
Containers@lists.linux-foundation.org
<https://lists.linux-foundation.org/mailman/listinfo/containers>
