**Date** Wed, 8 Sep 2021 11:38:56 -0400
**From** Steven Rostedt <>
**Subject** Re: WARNING in __static_key_slow_dec_deferred

On Wed, 8 Sep 2021 16:10:17 +0800
Hao Sun <sunhao.th@gmail.com> wrote:

> Hello,
>
> When using Healer to fuzz the latest Linux kernel, the following crash
> was triggered.

Thanks for the report. I think I have an idea of what happened.

>
> HEAD commit: ac08b1c68d1b Merge tag 'pci-v5.15-changes'
> git tree: upstream
> console output:
> https://drive.google.com/file/d/18VO37lgTvr6OLFJy4r0lOguCYXTV7mKD/view?usp=sharing
> kernel config: https://drive.google.com/file/d/1qrJUXD8ZIeAkg-xojzDpp04v9MtQ8RR6/view?usp=sharing
> C reproducer: https://drive.google.com/file/d/1JiIEm2ewm7fgdiHFXcODDGAtV70PYCXY/view?usp=sharing
> Syzlang reproducer:
> https://drive.google.com/file/d/1PjiSFUWZyLo655E8bTVfytsTjlMTN45F/view?usp=sharing
>
> If you fix this issue, please add the following tag to the commit:
> Reported-by: Hao Sun <sunhao.th@gmail.com>
>
> FAULT_INJECTION: forcing a failure.
> name failslab, interval 1, probability 0, space 0, times 0
> CPU: 2 PID: 9142 Comm: syz-executor Not tainted 5.14.0+ #15
> Hardware name: QEMU Standard PC (i440FX + PIIX, 1996), BIOS
> rel-1.12.0-59-gc9ba5276e321-prebuilt.qemu.org 04/01/2014
> Call Trace:
>  __dump_stack lib/dump_stack.c:88 [inline]
>  dump_stack_lvl+0x8d/0xcf lib/dump_stack.c:105
>  fail_dump lib/fault-inject.c:52 [inline]
>  should_fail+0x13c/0x160 lib/fault-inject.c:146
>  should_failslab+0x5/0x10 mm/slab_common.c:1326
>  slab_pre_alloc_hook.constprop.100+0x4e/0xc0 mm/slab.h:494
>  slab_alloc_node mm/slub.c:2880 [inline]
>  slab_alloc mm/slub.c:2967 [inline]
>  kmem_cache_alloc+0x44/0x2a0 mm/slub.c:2972
>  kmem_cache_zalloc include/linux/slab.h:711 [inline]
>  lsm_file_alloc security/security.c:572 [inline]
>  security_file_alloc+0x2c/0xb0 security/security.c:1515
>  __alloc_file+0x7f/0x150 fs/file_table.c:106
>  alloc_empty_file+0x4b/0x100 fs/file_table.c:150
>  alloc_file+0x31/0x170 fs/file_table.c:192
>  alloc_file_pseudo+0xb6/0x120 fs/file_table.c:232
>  __shmem_file_setup.part.53+0xb9/0x150 mm/shmem.c:4085
>  __shmem_file_setup mm/shmem.c:4148 [inline]
>  shmem_kernel_file_setup mm/shmem.c:4104 [inline]
>  shmem_zero_setup+0x6b/0x1f0 mm/shmem.c:4148
>  mmap_region+0x62e/0x790 mm/mmap.c:1824
>  do_mmap+0x438/0x670 mm/mmap.c:1575
>  vm_mmap_pgoff+0x10d/0x1b0 mm/util.c:519
>  vm_mmap+0x60/0x80 mm/util.c:538
>  __x86_set_memory_region+0x233/0x340 arch/x86/kvm/x86.c:11271
>  alloc_apic_access_page arch/x86/kvm/vmx/vmx.c:3648 [inline]
>  vmx_create_vcpu+0xc4b/0x1930 arch/x86/kvm/vmx/vmx.c:6871

The force failed allocation happened while creating a vcpu.

>  kvm_arch_vcpu_create+0x256/0x460 arch/x86/kvm/x86.c:10724
>  kvm_vm_ioctl_create_vcpu
> arch/x86/kvm/../../../virt/kvm/kvm_main.c:3592 [inline]
>  kvm_vm_ioctl+0x57c/0x1180 arch/x86/kvm/../../../virt/kvm/kvm_main.c:4314
>  vfs_ioctl fs/ioctl.c:51 [inline]
>  __do_sys_ioctl fs/ioctl.c:874 [inline]
>  __se_sys_ioctl fs/ioctl.c:860 [inline]
>  __x64_sys_ioctl+0xb6/0x100 fs/ioctl.c:860
>  do_syscall_x64 arch/x86/entry/common.c:50 [inline]
>  do_syscall_64+0x34/0xb0 arch/x86/entry/common.c:80
>  entry_SYSCALL_64_after_hwframe+0x44/0xae
> RIP: 0033:0x46a9a9
> Code: f7 d8 64 89 02 b8 ff ff ff ff c3 66 0f 1f 44 00 00 48 89 f8 48
> 89 f7 48 89 d6 48 89 ca 4d 89 c2 4d 89 c8 4c 8b 4c 24 08 0f 05 <48> 3d
> 01 f0 ff ff 73 01 c3 48 c7 c1 bc ff ff ff f7 d8 64 89 01 48
> RSP: 002b:00007fad048d2c58 EFLAGS: 00000246 ORIG_RAX: 0000000000000010
> RAX: ffffffffffffffda RBX: 0000000000078c0a RCX: 000000000046a9a9
> RDX: 0000000000000000 RSI: 000000000000ae41 RDI: 0000000000000004
> RBP: 00007fad048d2c90 R08: 0000000000000000 R09: 0000000000000000
> R10: 0000000000000000 R11: 0000000000000246 R12: 0000000000000017
> R13: 0000000000000000 R14: 0000000000078c0a R15: 00007fff737be610
> ------------[ cut here ]------------
> jump label: negative count!
> WARNING: CPU: 2 PID: 9142 at kernel/jump_label.c:235
> static_key_slow_try_dec+0x88/0xa0 kernel/jump_label.c:235
> Modules linked in:
> CPU: 2 PID: 9142 Comm: syz-executor Not tainted 5.14.0+ #15
> Hardware name: QEMU Standard PC (i440FX + PIIX, 1996), BIOS
> rel-1.12.0-59-gc9ba5276e321-prebuilt.qemu.org 04/01/2014
> RSP: 0018:ffffc900027b3d90 EFLAGS: 00010282
> RAX: 0000000000000000 RBX: 00000000ffffffff RCX: ffffc90000b35000
> RDX: 0000000000040000 RSI: ffffffff812d185c RDI: 00000000ffffffff
> RBP: ffffffff860d83a0 R08: 0000000000000000 R09: 0000000000000001
> R10: 0000000000000000 R11: 0000000000000000 R12: 00000000ffffffff
> R13: ffff888018411040 R14: ffffc900027b71e8 R15: 0000000000000004
> FS:  00007fad048d3700(0000) GS:ffff88813dd00000(0000) knlGS:0000000000000000
> CS:  0010 DS: 0000 ES: 0000 CR0: 0000000080050033
> CR2: 0000563eb1f86d08 CR3: 0000000108bac000 CR4: 0000000000752ee0
> DR0: 0000000000000000 DR1: 0000000000000000 DR2: 0000000000000000
> DR3: 0000000000000000 DR6: 00000000fffe0ff0 DR7: 0000000000000400
> PKRU: 55555554
> Call Trace:
>  __static_key_slow_dec_deferred+0x28/0x70 kernel/jump_label.c:286
>  kvm_free_lapic+0xaf/0xd0 arch/x86/kvm/lapic.c:2211
>  kvm_arch_vcpu_create+0x2f7/0x460 arch/x86/kvm/x86.c:10751

The failed allocation was detected, and the error path was taken.

The above is here:

```
void kvm_free_lapic(struct kvm_vcpu *vcpu)
{
        struct kvm_lapic *apic = vcpu->arch.apic;

        if (!vcpu->arch.apic)
                return;

        hrtimer_cancel(&apic->lapic_timer.timer);

        if (!(vcpu->arch.apic_base & MSR_IA32_APICBASE_ENABLE))
                static_branch_slow_dec_deferred(&apic_hw_disabled); <<<----- bad jump label accounting

        if (!apic->sw_enabled)
                static_branch_slow_dec_deferred(&apic_sw_disabled);

        if (apic->regs)
                free_page((unsigned long)apic->regs);
```

```
        kfree(apic);
}

What likely happened, was that the error was taken before the
apic_hw_disabled jump label was set, but I'm guessing that the
apic_base is initialized before the error, making the above think that
it needs to "undo" something that was never done.

I'll let someone else look into the kvm code to figure out how this
happened.

-- Steve

>  kvm_vm_ioctl_create_vcpu
> arch/x86/kvm/../../../virt/kvm/kvm_main.c:3592 [inline]
>  kvm_vm_ioctl+0x57c/0x1180 arch/x86/kvm/../../../virt/kvm/kvm_main.c:4314
>  vfs_ioctl fs/ioctl.c:51 [inline]
>  __do_sys_ioctl fs/ioctl.c:874 [inline]
>  __se_sys_ioctl fs/ioctl.c:860 [inline]
>  __x64_sys_ioctl+0xb6/0x100 fs/ioctl.c:860
>  do_syscall_x64 arch/x86/entry/common.c:50 [inline]
>  do_syscall_64+0x34/0xb0 arch/x86/entry/common.c:80
>  entry_SYSCALL_64_after_hwframe+0x44/0xae
> RIP: 0033:0x46a9a9
> Code: f7 d8 64 89 02 b8 ff ff ff ff c3 66 0f 1f 44 00 00 48 89 f8 48
> 89 f7 48 89 d6 48 89 ca 4d 89 c2 4d 89 c8 4c 8b 4c 24 08 0f 05 <48> 3d
> 01 f0 ff ff 73 01 c3 48 c7 c1 bc ff ff ff f7 d8 64 89 01 48
> RSP: 002b:00007fad048d2c58 EFLAGS: 00000246 ORIG_RAX: 0000000000000010
> RAX: ffffffffffffffda RBX: 000000000078c0a0 RCX: 000000000046a9a9
> RDX: 0000000000000000 RSI: 000000000000ae41 RDI: 0000000000000004
> RBP: 00007fad048d2c90 R08: 0000000000000000 R09: 0000000000000000
> R10: 0000000000000000 R11: 0000000000000246 R12: 0000000000000017
> R13: 0000000000000000 R14: 000000000078c0a0 R15: 00007fff737be610%
```