

excessive TCP duplicate acks revisited

Source: <http://unix.derkeiler.com/Mailing-Lists/FreeBSD/current/2007-11/msg00582.html>

- *From:* Gregory Wright <gwright@xxxxxxxxxxx>
 - *Date:* Fri, 9 Nov 2007 13:07:39 -0500
-

(Note: long message)

Hi,

The tcp duplicate ACK attack is back.

Last March, there was a thread on duplicate TCP acks in -CURRENT. I have been able to reproduce the problem on 7.0-BETA2 (amd64) and have some new information that might help locate the bug.

Background: I first noticed problems with tcp connections dropping when Bacula was running on our backup server. The hardware was a dual Opteron 244, 2 GB RAM, running FreeBSD 6.2-RELEASE-p2. The ethernet NIC was a bge.

We went through an extensive process to rule out hardware and cabling problems: the server hardware was replaced with a single Opteron 270 (dual core) on a Tyan S2882-D motherboard. The memory was replaced as well. The NICs are still bge. This machine is "hardtack".

We have another server in the same rack, a dual Opteron 2214, 4 GB RAM, running FreeBSD 6.2-RELEASE-p1. It has never shown the "dropped connection" problem. This machine is "greenhouse-george"

All of the machines are on a single LAN, with a Nortel BayStack 450-24T ethernet switch.

Experiments:

The first experiment was to use netperf to send a tcp stream from either a PowerBook G4 (OS X 10.4.10) or a dual Intel Clovertown box running FreeBSD 6.2-RELEASE-p8. This is from the PowerBook:

```
crossroads-able> netperf -H greenhouse-george
TCP STREAM TEST from localhost (0.0.0.0) port 0 AF_INET to greenhouse- george.18clay.com
(192.168.2.63) port 0 AF_INET
Recv Send Send
Socket Socket Message Elapsed
Size Size Size Time Throughput
```

excessive TCP duplicate acks revisited

bytes bytes bytes secs. 10^6bits/sec

65536 262144 262144 10.02 93.77

crossroads-able> netperf -H hardtack

TCP STREAM TEST from localhost (0.0.0.0) port 0 AF_INET to hardtack. 18clay.com (192.168.2.61) port 0 AF_INET

Recv Send Send

Socket Socket Message Elapsed

Size Size Size Time Throughput

bytes bytes bytes secs. 10^6bits/sec

65536 262144 262144 10.10 29.24

crossroads-able>

(That there is a problem in the second case is easy to see just from the ethernet switch: the activity lights blink on and off. A wireshark trace shows a duplicate ACK storm and gaps in the transmitted packet stream.)

Just to be certain that this isn't a weird cabling problem, I swapped the cables to the two machines:

crossroads-able> netperf -H greenhouse-george

TCP STREAM TEST from localhost (0.0.0.0) port 0 AF_INET to greenhouse- george.18clay.com (192.168.2.63) port 0 AF_INET

Recv Send Send

Socket Socket Message Elapsed

Size Size Size Time Throughput

bytes bytes bytes secs. 10^6bits/sec

65536 262144 262144 10.02 92.68

crossroads-able> netperf -H hardtack

TCP STREAM TEST from localhost (0.0.0.0) port 0 AF_INET to hardtack. 18clay.com (192.168.2.61) port 0 AF_INET

Recv Send Send

Socket Socket Message Elapsed

Size Size Size Time Throughput

bytes bytes bytes secs. 10^6bits/sec

65536 262144 262144 10.02 30.86

crossroads-able>

The low throughput (and gaps in the tcp stream) stay with the machine rather than moving with the cable.

When I use my dual Clovertown box (FreeBSD 6.2-RELEASE-p8, em NIC), as a source, I get

ivy-mike# netperf -H greenhouse-george

TCP STREAM TEST from 0.0.0.0 (0.0.0.0) port 0 AF_INET to greenhouse- george.18clay.com (192.168.2.63) port 0 AF_INET

excessive TCP duplicate acks revisited

excessive TCP duplicate acks revisited

```
Recv Send Send
Socket Socket Message Elapsed
Size Size Size Time Throughput
bytes bytes bytes secs. 10^6bits/sec
```

```
65536 32768 32768 10.58 94.14
```

```
ivy-mike# netperf -H hardtack
```

```
TCP STREAM TEST from 0.0.0.0 (0.0.0.0) port 0 AF_INET to hardtack. 18clay.com (192.168.2.61) port 0
AF_INET
```

```
Recv Send Send
Socket Socket Message Elapsed
Size Size Size Time Throughput
bytes bytes bytes secs. 10^6bits/sec
```

```
65536 32768 32768 10.57 89.15
```

```
ivy-mike#
```

FreeBSD seems not as disturbed by the extra ACKs as OS X, but I still see the activity lights on the switch blinking, indicating dropouts in the tcp stream. Also, longer tests using either source machine almost always result in failure ("Broken Pipe") after a few minutes.

Now is where things get interesting. If I connect the PowerBook to hardtack (the machine showing the bug) with just a cable, bypassing the switch, I get

```
crossroads-able> netperf -H 192.168.2.2 -l 300
```

```
TCP STREAM TEST from (null) (0.0.0.0) port 0 AF_INET to (null) (192.168.2.2) port 0 AF_INET
```

```
Recv Send Send
Socket Socket Message Elapsed
Size Size Size Time Throughput
bytes bytes bytes secs. 10^6bits/sec
```

```
65536 262144 262144 300.54 330.45
```

```
crossroads-able>
```

The higher throughput is because I have two GbE interfaces directly connected, instead of going through the 10/100 Mb switch. The key thing is that I can run this test for hours without the connection dropping. (I have run it for as long as 10000 seconds.) And the wireshark log shows no duplicate ACK storm.

Is there perhaps something wrong with the switch? I swapped out our BayStack 450-24T for an identical unit with the same results. I also have an unmanaged SMC EZ Switch 10/100, so I tried a minimal setup:

```
PowerBook <-----> SMC switch <-----> FreeBSD 7.0-BETA2 server (hardtack)
```

excessive TCP duplicate acks revisited

Only the PowerBook and FreeBSD server are connected to the switch. A netperf tcp streaming test gives:

```
crossroads-able> netperf -H 192.168.2.2
TCP STREAM TEST from (null) (0.0.0.0) port 0 AF_INET to (null) (192.168.2.2) port 0 AF_INET
Recv Send Send
Socket Socket Message Elapsed
Size Size Size Time Throughput
bytes bytes bytes secs. 10^6bits/sec

65536 262144 262144 10.02 24.25
crossroads-able>
```

The throughput is low, but most importantly, the wireshark log shows that the duplicate ACK storm is back.

So is there something coming out of the switch that FreeBSD's network stack doesn't like? No, this doesn't seem to be the case. The old SMC switch doesn't even send out spanning tree packets.

Next experiment: netperf tcp stream from the PowerBook, but with the interface forced into 100TX full duplex. Result:

```
crossroads-able> netperf -H 192.168.2.2
```

```
TCP STREAM TEST from (null) (0.0.0.0) port 0 AF_INET to (null) (192.168.2.2) port 0 AF_INET
Recv Send Send
Socket Socket Message Elapsed
Size Size Size Time Throughput
bytes bytes bytes secs. 10^6bits/sec

65536 262144 262144 10.02 33.81
crossroads-able>
```

So the throughput has dropped and the wireshark log shows duplicate ACK storms again.

If I force the PowerBook to 100TX half duplex the wireshark log still shows duplicate ACK storms. And at 10 Mb/s, full or half duplex, I see still bursts of duplicate ACKs.

One last thing to note is that I can run netperf tests from the buggy machine to the laptop with no trouble. Again forcing the PowerBook interface to 100TX full duplex,

```
TCP STREAM TEST from 0.0.0.0 (0.0.0.0) port 0 AF_INET to 192.168.2.1 (192.168.2.1) port 0 AF_INET
Recv Send Send
Socket Socket Message Elapsed
Size Size Size Time Throughput
bytes bytes bytes secs. 10^6bits/sec

262144 32768 32768 10.02 93.03
```

excessive TCP duplicate acks revisited

excessive TCP duplicate acks revisited

hardtack#

The wireshark logs show no duplicate ACKs.

One interesting thing is that the packet logs for the reverse (properly working) case show a more regular structure of data and ACK packets. In the forward (buggy) direction, the data packets are burstier, and to my eye more irregular. This was mentioned in passing in the original thread, in which someone speculated that there might be a window problem.

So, in sum:

The duplicate ack problem happens on my FreeBSD 7.0-BETA2 machine (single Opteron 270, Tyan S2882-D, 2 GB RAM) when the bge interface is operating in 10 or 100 Mb/s mode, but not in Gb mode.

It does not happen on my FreeBSD 6.2-RELEASE-p1 box (dual Opteron 2214, Tyan 3992G3NR, 4 GB RAM) which has a bge NIC as well. Nor does it happen on my dual Clovertown box (dual E5345, Tyan S2696A2NRF, 4 GB) with em NIC.

The buggy machine is not much use to me until this problem can be fixed, so I can dedicate it to debugging for a while. Any hints about where code might be instrumented to track the bug down or new experiments to perform are welcome.

I can provide wireshark logs; compressed they are not too large (a few MB).

Here is the dmesg of the buggy machine:

```
hardtack# dmesg
Copyright (c) 1992-2007 The FreeBSD Project.
Copyright (c) 1979, 1980, 1983, 1986, 1988, 1989, 1991, 1992, 1993, 1994
The Regents of the University of California. All rights reserved.
FreeBSD is a registered trademark of The FreeBSD Foundation.
FreeBSD 7.0-BETA2 #0: Fri Nov 2 14:54:38 UTC 2007
root@xxxxxxxxxxxxxxxxxxxxxx:/usr/obj/usr/src/sys/GENERIC
Timecounter "i8254" frequency 1193182 Hz quality 0
CPU: Dual Core AMD Opteron(tm) Processor 270 HE (1993.41-MHz K8-class CPU)
Origin = "AuthenticAMD" Id = 0x20f12 Stepping = 2
Features=0x178bfbff<FPU,VME,DE,PSE,TSC,MSR,PAE,MCE,CX8,APIC,SEP,MTRR,PGE
,MCA,CMOV,PAT,PSE36,CLFLUSH,MMX,FXSR,SSE,SSE2,HTT>
Features2=0x1<SSE3>
AMD Features=0xe2500800<SYSCALL,NX,MMX+,FFXSR,LM,3DNow!+,3DNow!>
AMD Features2=0x3<LAHF,CMP>
Cores per package: 2
usable memory = 2134962176 (2036 MB)
```

excessive TCP duplicate acks revisited

excessive TCP duplicate acks revisited

avail memory = 2060251136 (1964 MB)
ACPI APIC Table: <A M I OEMAPIC >
FreeBSD/SMP: Multiprocessor System Detected: 2 CPUs
cpu0 (BSP): APIC ID: 0
cpu1 (AP): APIC ID: 1
MADT: Forcing active-low polarity and level trigger for SCI
ioapic0 <Version 1.1> irqs 0-23 on motherboard
ioapic1 <Version 1.1> irqs 24-27 on motherboard
ioapic2 <Version 1.1> irqs 28-31 on motherboard
kbd1 at kbdmux0
ath_hal: 0.9.20.3 (AR5210, AR5211, AR5212, RF5111, RF5112, RF2413, RF5413)
acpi0: <A M I OEMXSDT> on motherboard
acpi0: [ITHREAD]
acpi0: Power Button (fixed)
acpi0: reservation of 0, a0000 (3) failed
acpi0: reservation of 100000, 7ff00000 (3) failed
Timecounter "ACPI-fast" frequency 3579545 Hz quality 1000
acpi_timer0: <24-bit timer at 3.579545MHz> port 0x1008-0x100b on acpi0
acpi_hpet0: <High Precision Event Timer> iomem 0xfec01000-0xfec013ff on acpi0
Timecounter "HPET" frequency 14318180 Hz quality 900
cpu0: <ACPI CPU> on acpi0
acpi_throttle0: <ACPI CPU Throttling> on cpu0
powernow0: <Cool`n`Quiet K8> on cpu0
cpu1: <ACPI CPU> on acpi0
powernow1: <Cool`n`Quiet K8> on cpu1
pcib0: <ACPI Host-PCI bridge> port 0xcf8-0xcff on acpi0
pci0: <ACPI PCI bus> on pcib0
pci1: <ACPI PCI-PCI bridge> at device 6.0 on pci0
pci3: <ACPI PCI bus> on pcib1
ohci0: <OHCI (generic) USB controller> mem 0xfeafb000-0xfeafbfff irq 19 at device 0.0 on pci3
ohci0: [GIANT-LOCKED]
ohci0: [ITHREAD]
usb0: OHCI version 1.0, legacy support
usb0: <OHCI (generic) USB controller> on ohci0
usb0: USB revision 1.0
uhub0: <AMD OHCI root hub, class 9/0, rev 1.00/1.00, addr 1> on usb0
uhub0: 3 ports with 3 removable, self powered
ohci1: <OHCI (generic) USB controller> mem 0xfeafc000-0xfeafcfff irq 19 at device 0.1 on pci3
ohci1: [GIANT-LOCKED]
ohci1: [ITHREAD]
usb1: OHCI version 1.0, legacy support
usb1: <OHCI (generic) USB controller> on ohci1
usb1: USB revision 1.0
uhub1: <AMD OHCI root hub, class 9/0, rev 1.00/1.00, addr 1> on usb1
uhub1: 3 ports with 3 removable, self powered
atapci0: <SiI 3114 SATA150 controller> port
0xbc00-0xbc07,0xb880-0xb883,0xb800-0xb807,0xac00-0xac03,0xa880-0xa88f mem
0xfeafec00-0xfeafefff irq 19 at device 5.0 on pci3
atapci0: [ITHREAD]
ata2: <ATA channel 0> on atapci0
ata2: [ITHREAD]

excessive TCP duplicate acks revisited

ata3: <ATA channel 1> on atapci0
ata3: [ITHREAD]
ata4: <ATA channel 2> on atapci0
ata4: [ITHREAD]
ata5: <ATA channel 3> on atapci0
ata5: [ITHREAD]
vgapci0: <VGA-compatible display> port 0xb000-0xb0ff mem
0xfd000000-0xfdffffff,0xfeaff000-0xfeafffff irq 18 at device 6.0 on pci3
fxp0: <Intel 82551 Pro/100 Ethernet> port 0xa800-0xa83f mem
0xfeafa000-0xfeafafff,0xfeaa0000-0xfeabffff irq 18 at device 8.0 on pci3
miibus0: <MII bus> on fxp0
inphy0: <i82555 10/100 media interface> PHY 1 on miibus0
inphy0: 10baseT, 10baseT-FDX, 100baseTX, 100baseTX-FDX, auto
fxp0: Ethernet address: 00:e0:81:4b:3e:39
fxp0: [ITHREAD]
isab0: <PCI-ISA bridge> at device 7.0 on pci0
isa0: <ISA bus> on isab0
atapci1: <AMD 8111 UDMA133 controller> port 0x1f0-0x1f7,0x3f6,0x170-0x177,0x376,0xffa0-0xffaf at
device 7.1 on pci0
ata0: <ATA channel 0> on atapci1
ata0: [ITHREAD]
ata1: <ATA channel 1> on atapci1
ata1: [ITHREAD]
pci0: <serial bus, SMBus> at device 7.2 (no driver attached)
pci0: <bridge> at device 7.3 (no driver attached)
pcib2: <ACPI PCI-PCI bridge> at device 10.0 on pci0
pci2: <ACPI PCI bus> on pcib2
ahc0: <Adaptec 3960D Ultra160 SCSI adapter> port 0x8400-0x84ff mem 0xfc8fd000-0xfc8fdfff irq 27 at
device 3.0 on pci2
ahc0: [ITHREAD]
aic7899: Ultra160 Wide Channel A, SCSI Id=7, 32/253 SCBs
ahc1: <Adaptec 3960D Ultra160 SCSI adapter> port 0x8800-0x88ff mem 0xfc8fe000-0xfc8fefff irq 24 at
device 3.1 on pci2
ahc1: [ITHREAD]
aic7899: Ultra160 Wide Channel B, SCSI Id=7, 32/253 SCBs
pci0:2:9:0: bad VPD cksum, remain 72
bge0: <Broadcom Gigabit Ethernet Controller, ASIC rev. 0x2100> mem 0xfc870000-0xfc87ffff irq 24 at
device 9.0 on pci2
miibus1: <MII bus> on bge0
brgphy0: <BCM5704 10/100/1000baseTX PHY> PHY 1 on miibus1
brgphy0: 10baseT, 10baseT-FDX, 100baseTX, 100baseTX-FDX, 1000baseT, 1000baseT-FDX, auto
bge0: Ethernet address: 00:e0:81:4b:3e:ae
bge0: [ITHREAD]
pci0:2:9:1: bad VPD cksum, remain 72
bge1: <Broadcom Gigabit Ethernet Controller, ASIC rev. 0x2100> mem 0xfc890000-0xfc89ffff irq 25 at
device 9.1 on pci2
miibus2: <MII bus> on bge1
brgphy1: <BCM5704 10/100/1000baseTX PHY> PHY 1 on miibus2
brgphy1: 10baseT, 10baseT-FDX, 100baseTX, 100baseTX-FDX, 1000baseT, 1000baseT-FDX, auto
bge1: Ethernet address: 00:e0:81:4b:3e:af
bge1: [ITHREAD]

excessive TCP duplicate acks revisited

pcib3: <ACPI PCI-PCI bridge> at device 11.0 on pci0
pci1: <ACPI PCI bus> on pcib3
acpi_button0: <Power Button> on acpi0
atkbd0: <Keyboard controller (i8042)> port 0x60,0x64 irq 1 on acpi0
atkbd0: <AT Keyboard> irq 1 on atkbd0
kbd0 at atkbd0
atkbd0: [GIANT-LOCKED]
atkbd0: [ITHREAD]
psm0: <PS/2 Mouse> irq 12 on atkbd0
psm0: [GIANT-LOCKED]
psm0: [ITHREAD]
psm0: model IntelliMouse, device ID 3
sio0: configured irq 4 not in bitmap of probed irqs 0
sio0: port may not be enabled
sio0: configured irq 4 not in bitmap of probed irqs 0
sio0: port may not be enabled
sio0: <16550A-compatible COM port> port 0x3f8-0x3ff irq 4 flags 0x10 on acpi0
sio0: type 16550A
sio0: [FILTER]
sio1: configured irq 3 not in bitmap of probed irqs 0
sio1: port may not be enabled
sio1: configured irq 3 not in bitmap of probed irqs 0
sio1: port may not be enabled
sio1: <16550A-compatible COM port> port 0x2f8-0x2ff irq 3 on acpi0
sio1: type 16550A
sio1: [FILTER]
fdc0: <floppy drive controller (FDE)> port 0x3f0-0x3f5,0x3f7 irq 6 drq 2 on acpi0
fdc0: [FILTER]
fd0: <1440-KB 3.5" drive> on fdc0 drive 0
orm0: <ISA Option ROMs> at iomem 0xc0000-0xc7fff,0xc8000-0xcc7ff on isa0
ppc0: cannot reserve I/O port range
sc0: <System console> at flags 0x100 on isa0
sc0: VGA <16 virtual consoles, flags=0x300>
vga0: <Generic ISA VGA> at port 0x3c0-0x3df iomem 0xa0000-0xbffff on isa0
Timecounters tick every 1.000 msec
acd0: DVDROM <ATAPI DVD D DH16D2P/HP56> at ata1-master UDMA33
ad4: 305245MB <Seagate ST3320620AS 3.AAE> at ata2-master SATA150
Waiting 5 seconds for SCSI devices to settle
sa0 at ahc1 bus 0 target 6 lun 0
sa0: <IBM ULTRIUM-TD2 3AYD> Removable Sequential Access SCSI-3 device
sa0: 160.000MB/s transfers (80.000MHz DT, offset 31, 16bit)
SMP: AP CPU #1 Launched!
Trying to mount root from ufs:/dev/ad4s1a
WARNING: / was not properly dismounted
ch0 at ahc1 bus 0 target 5 lun 0
ch0: <DELL PV-122T K17r> Removable Changer SCSI-2 device
ch0: 3.300MB/s transfers
ch0: 8 slots, 1 drive, 1 picker, 0 portals
bge1: link state changed to DOWN
bge1: link state changed to UP
bge1: promiscuous mode enabled

excessive TCP duplicate acks revisited

bge1: promiscuous mode disabled
bge1: link state changed to DOWN
bge1: link state changed to UP
bge1: link state changed to DOWN
bge1: link state changed to UP
bge1: link state changed to DOWN
bge1: link state changed to UP
bge1: link state changed to DOWN
bge1: link state changed to UP
bge1: link state changed to DOWN
bge1: link state changed to UP
bge1: link state changed to DOWN
bge1: link state changed to UP
hardtack#

Please let me know what I can do to track this down.

Best Wishes,
Greg

Gregory Wright
Antiope Associates LLC
18 Clay Street
Fair Haven, New Jersey 07704
USA

1 (732) 924-4549
1 (732) 345-8378 [fax]

freebsd-current@xxxxxxxxxxx mailing list
<http://lists.freebsd.org/mailman/listinfo/freebsd-current>
To unsubscribe, send any mail to "freebsd-current-unsubscribe@xxxxxxxxxxx"