Subject: seems to be a flaw in cfq
Posted by Vasily Tarasov on Tue, 19 Dec 2006 13:53:04 GMT
View Forum Message <> Reply to Message

ello, Jens.

Seems, that we've found some problem in CFQ.
I used your fio tool of version 1.9 to reproduce it.
2.6.18 vanilla kernel.
Generally situation is the following:
there are several readers, which read from their own files, except
two readers, which read one shared file.
These two readers experienced huge starvation!

I want to note, that there is no such situation on deadline scheduler
and situation is much better for anticipatory scheduler.

Here is a job file:

<snip>
; cfq bug reproduce
; each file is 1Gib size

[global]
timeout=200

[reader1]
filename=file1

[reader2]
filename=file2

[reader3]
filename=file3

; --- this two reads the same file ---
[reader4]
filename=file4

[reader5]
filename=file4
; -----------------------------------

[reader6]
filename=file6

[reader7]
filename=file7

[reader8]
filename=file8

[reader9]
filename=file9

[reader10]
filename=file10

[reader11]
filename=file11
</snip>

Here is an output of fio:
fio job1.file
reader1: (g=0): rw=read, odir=1, bs=4K-4K/4K-4K, rate=0, ioengine=sync,
iodepth=1
reader2: (g=0): rw=read, odir=1, bs=4K-4K/4K-4K, rate=0, ioengine=sync,
iodepth=1
reader3: (g=0): rw=read, odir=1, bs=4K-4K/4K-4K, rate=0, ioengine=sync,
iodepth=1
reader4: (g=0): rw=read, odir=1, bs=4K-4K/4K-4K, rate=0, ioengine=sync,
iodepth=1
reader5: (g=0): rw=read, odir=1, bs=4K-4K/4K-4K, rate=0, ioengine=sync,
iodepth=1
reader6: (g=0): rw=read, odir=1, bs=4K-4K/4K-4K, rate=0, ioengine=sync,
iodepth=1
reader7: (g=0): rw=read, odir=1, bs=4K-4K/4K-4K, rate=0, ioengine=sync,
iodepth=1
reader8: (g=0): rw=read, odir=1, bs=4K-4K/4K-4K, rate=0, ioengine=sync,
iodepth=1
reader9: (g=0): rw=read, odir=1, bs=4K-4K/4K-4K, rate=0, ioengine=sync,
iodepth=1
reader10: (g=0): rw=read, odir=1, bs=4K-4K/4K-4K, rate=0, ioengine=sync,
iodepth=1
reader11: (g=0): rw=read, odir=1, bs=4K-4K/4K-4K, rate=0, ioengine=sync,
iodepth=1
Starting 11 threads
Threads running: 11: [RRRRRRRRRRR] [100.00% done] [31298/ 0 kb/s]
[eta 00m:00s]
reader1: (groupid=0): err= 0:
read : io= 662MiB, bw= 3469KiB/s, runt=200058msec
slat (msec): min= 0, max= 0, avg= 0.00, dev= 0.00
clat (msec): min= 0, max= 1362, avg= 1.06, dev=30.90

bw (KiB/s) : min= 490, max=12582, per=12.03%, avg=3732.10, dev=4034.82
cpu : usr=0.15%, sys=4.43%, ctx=169761
reader2: (groupid=0): err= 0:
read : io= 644MiB, bw= 3375KiB/s, runt=200041msec
slat (msec): min= 0, max= 0, avg= 0.00, dev= 0.00
clat (msec): min= 0, max= 1454, avg= 1.09, dev=31.69
bw (KiB/s) : min= 485, max=12558, per=11.59%, avg=3595.80, dev=3795.64
cpu : usr=0.15%, sys=4.36%, ctx=165149
reader3: (groupid=0): err= 0:
read : io= 679MiB, bw= 3561KiB/s, runt=200057msec
slat (msec): min= 0, max= 0, avg= 0.00, dev= 0.00
clat (msec): min= 0, max= 1584, avg= 1.03, dev=30.33
bw (KiB/s) : min= 445, max=12689, per=12.53%, avg=3886.43, dev=4161.05
cpu : usr=0.15%, sys=4.63%, ctx=174219
reader4: (groupid=0): err= 0:
read : io= 1MiB, bw= 9KiB/s, runt=200009msec
<<< ONLY ONE MIB
slat (msec): min= 0, max= 0, avg= 0.00, dev= 0.00
clat (msec): min= 0, max= 1327, avg=411.34, dev=597.62
bw (KiB/s) : min= 3, max= 24, per=0.03%, avg= 9.92, dev=11.60
cpu : usr=0.00%, sys=0.01%, ctx=501
reader5: (groupid=0): err= 0:
read : io= 1MiB, bw= 9KiB/s, runt=200009msec
<<< ONLY ONE MIB
slat (msec): min= 0, max= 0, avg= 0.00, dev= 0.00
clat (msec): min= 0, max= 1327, avg=413.70, dev=599.16
bw (KiB/s) : min= 3, max= 24, per=0.03%, avg= 9.67, dev=11.18
cpu : usr=0.00%, sys=0.01%, ctx=491
reader6: (groupid=0): err= 0:
read : io= 661MiB, bw= 3466KiB/s, runt=200045msec
slat (msec): min= 0, max= 0, avg= 0.00, dev= 0.00
clat (msec): min= 0, max= 1453, avg= 1.06, dev=30.92
bw (KiB/s) : min= 483, max=12222, per=12.14%, avg=3765.32, dev=4002.38
cpu : usr=0.14%, sys=4.51%, ctx=169566
reader7: (groupid=0): err= 0:
read : io= 655MiB, bw= 3435KiB/s, runt=200056msec
slat (msec): min= 0, max= 0, avg= 0.00, dev= 0.00
clat (msec): min= 0, max= 1585, avg= 1.07, dev=31.36
bw (KiB/s) : min= 602, max=12115, per=11.78%, avg=3654.40, dev=3858.70
cpu : usr=0.14%, sys=4.48%, ctx=168098
reader8: (groupid=0): err= 0:
read : io= 666MiB, bw= 3496KiB/s, runt=200000msec
slat (msec): min= 0, max= 0, avg= 0.00, dev= 0.00
clat (msec): min= 0, max= 1358, avg= 1.05, dev=30.95
bw (KiB/s) : min= 7, max=12730, per=12.23%, avg=3791.85, dev=4177.52
cpu : usr=0.15%, sys=4.53%, ctx=170999
reader9: (groupid=0): err= 0:
read : io= 666MiB, bw= 3495KiB/s, runt=200016msec

slat (msec): min= 0, max= 0, avg= 0.00, dev= 0.00
clat (msec): min= 0, max= 1456, avg= 1.05, dev=30.78
bw (KiB/s) : min= 6, max=12681, per=12.13%, avg=3761.05, dev=4038.50
cpu : usr=0.17%, sys=4.52%, ctx=170982
reader10: (groupid=0): err= 0:
read : io= 648MiB, bw= 3398KiB/s, runt=200026msec
slat (msec): min= 0, max= 0, avg= 0.00, dev= 0.00
clat (msec): min= 0, max= 1356, avg= 1.09, dev=31.57
bw (KiB/s) : min= 5, max=12533, per=11.79%, avg=3657.35, dev=3930.12
cpu : usr=0.15%, sys=4.35%, ctx=166233
reader11: (groupid=0): err= 0:
read : io= 628MiB, bw= 3296KiB/s, runt=200049msec
slat (msec): min= 0, max= 0, avg= 0.00, dev= 0.00
clat (msec): min= 0, max= 1591, avg= 1.12, dev=32.34
bw (KiB/s) : min= 3, max=13910, per=11.59%, avg=3595.44, dev=4000.10
cpu : usr=0.16%, sys=4.27%, ctx=161300

Run status group 0 (all jobs):
READ: io=5917MiB, aggrb=31013, minb=9, maxb=3561, mint=200000msec,
maxt=200058msec

Disk stats (read/write):
sda: ios=1516020/65, merge=242/28, ticks=1975258/49498,
in_queue=2024681, util=100.00%

Is such behavior expected?

Thanks in advance,
Vasily.