Ceph - Bug #752

High I/O wait when writing data

01/28/2011 01:17 PM - Wido den Hollander

Status:	Resolved	**Start date:**	01/28/2011
Priority:	Normal	**Due date:**	
Assignee:		**% Done:**	0%
Category:	OSD	**Estimated time:**	0.00 hour
Target version:		**Spent time:**	19.00 hours
Source:		**Reviewed:**	
Tags:		**Affected Versions:**	
Backport:		**ceph-qa-suite:**	
Regression:	No	**Pull request ID:**	
Severity:	3 - minor		

Description

Like I said on IRC, I'm seeing a high load on my machine "noisy".

The setup is:

- Intel(R) Xeon(R) CPU 5110 1.6Ghz (Dual-Core)
- 6GB DDR2
- 1x 80GB (OS / Logging)
- 4x WD 2TB (OSD data)

I have 4 OSD's running (on the 2TB disks), with their journal on a tmpfs drive (200MB per OSD).

With Qemu-RBD I have a VM "beta" running, this VM has a 1.5TB disk and I'm rsyncing 1.1TB data to it.

The replication level of the "rbd" pool is at 3:

```
pg_pool 3 'rbd' pg_pool(rep pg_size 3 crush_ruleset 3 object_hash rjenkins pg_num 2048 pgp_num 256
 lpg_num 2 lpgp_num 2 last_change 9 owner 0)
```

The status of my Ceph cluster:

```
2011-01-28 22:12:37.084838    pg v123162: 8224 pgs: 8219 active+clean, 5 active+clean+inconsistent
; 745 GB data, 2298 GB used, 5147 GB / 7452 GB avail
2011-01-28 22:12:37.104838   mds e66: 1/1/1 up {0=up:active}
2011-01-28 22:12:37.128499   osd e2064: 4 osds: 4 up, 4 in
2011-01-28 22:12:37.128602   log 2011-01-28 22:12:11.549246 osd1 [2a00:f10:113:1:230:48ff:fe8d:a21
e]:6804/2622 721 : [INF] 3.db scrub ok
2011-01-28 22:12:37.139256   class rbd (v1.3 [x86-64])
2011-01-28 22:12:37.147454   mon e1: 1 mons at {noisy=[2a00:f10:113:1:230:48ff:fe8d:a21e]:6789/0}
```

In the VM I'm downloading the data with about 100Mbit, but the speed fluctuates between 5MB and 9MB/sec.

The load on noisy is rather high, a sysstat tells me:

```
04:55:01 PM     CPU     %user     %nice   %system   %iowait    %steal     %idle
05:05:02 PM     all     26.10      0.00     42.90     24.24      0.00      6.76
05:15:01 PM     all     25.54      0.00     42.41     24.20      0.00      7.85
05:25:02 PM     all     25.75      0.00     42.64     23.59      0.00      8.02
05:35:01 PM     all     24.13      0.00     40.21     27.19      0.00      8.47
05:45:02 PM     all     24.81      0.00     41.38     26.04      0.00      7.77
05:55:02 PM     all     22.08      0.00     36.88     28.23      0.00     12.82
```

```
06:05:01 PM     all     23.39      0.00     38.34     27.97      0.00     10.31
06:15:02 PM     all     25.14      0.00     41.46     25.76      0.00      7.64
06:25:02 PM     all     25.22      0.00     41.45     25.28      0.00      8.05
06:35:02 PM     all     24.22      0.00     40.70     25.31      0.00      9.77
06:45:01 PM     all     25.08      0.00     41.44     25.11      0.00      8.37
06:55:02 PM     all     24.03      0.00     39.25     27.08      0.00      9.65
07:05:01 PM     all     24.48      0.00     39.72     25.35      0.00     10.46
07:15:01 PM     all     23.48      0.00     38.83     27.38      0.00     10.31
07:25:02 PM     all     22.52      0.00     37.76     29.40      0.00     10.32
07:35:01 PM     all     22.48      0.00     36.90     29.93      0.00     10.69
07:45:01 PM     all     21.72      0.00     35.86     31.07      0.00     11.35
07:55:02 PM     all     22.54      0.00     38.56     29.19      0.00      9.71
08:05:02 PM     all     23.96      0.00     40.21     27.77      0.00      8.06
08:15:02 PM     all     22.71      0.00     37.98     31.26      0.00      8.05
08:25:02 PM     all     20.86      0.00     34.94     30.19      0.00     14.00
08:35:01 PM     all     24.32      0.00     40.13     27.69      0.00      7.86
08:45:02 PM     all     25.04      0.00     41.97     27.34      0.00      5.64
08:55:01 PM     all     24.07      0.00     40.50     27.60      0.00      7.82
09:05:01 PM     all     23.60      0.00     39.41     29.10      0.00      7.90
09:15:01 PM     all     18.31      0.00     31.59     33.26      0.00     16.84
09:25:01 PM     all     24.57      0.00     41.64     28.17      0.00      5.62
09:35:02 PM     all     23.22      0.00     39.63     28.83      0.00      8.32
09:45:01 PM     all     20.53      0.00     35.36     34.28      0.00      9.83
09:55:02 PM     all     23.67      0.00     39.81     26.93      0.00      9.59
10:05:02 PM     all     23.44      0.00     40.71     28.88      0.00      6.97
Average:        all     24.71      0.00     39.26     24.85      0.00     11.19
```

As you can see, the I/O Wait is high, but the System also consumes a lot of CPU.

On Sage's request I set debug-ms to 1 for osd.0, the last 10.000 lines of the log are attached to this issue.

Is this a normal load? With replication at three 100Mb of traffic would generate 300Mb of writes spread over the 4 disks, that seems to be in the range of what these disks should be able to handle.

The rsync just finished and after it did, the load dropped very hard, the rsync stats:

```
Number of files: 932
Number of files transferred: 473
Total file size: 775568137247 bytes
Total transferred file size: 670012521514 bytes
Literal data: 670012521514 bytes
Matched data: 0 bytes
File list size: 51337
File list generation time: 0.199 seconds
File list transfer time: 0.000 seconds
Total bytes sent: 10432
Total bytes received: 670094382303

sent 10432 bytes  received 670094382303 bytes  7369464.94 bytes/sec
total size is 775568137247  speedup is 1.16
```

I had to restart the rsync a few times since the VM crashed due to some btrfs bugs (which are fixed now).

P.S.: I saw #563 a lot during this rsync.

History

#1 - 01/31/2011 11:22 AM - Wido den Hollander

I've done some benchmarks today in the VM and noticed something weird.

In the VM I ran:

```
dd if=/dev/zero of=1GB.bin bs=1024k count=1024 conv=sync
```

This gave me a really poor performance:

```
root@beta:~# dd if=/dev/zero of=1GB.bin bs=1024k count=1024 conv=sync && sync
1024+0 records in
1024+0 records out
1073741824 bytes (1.1 GB) copied, 57.4975 s, 18.7 MB/s
root@beta:~#
```

But what I noticed more, is the about of written data to the disks:

```
root@noisy:~# iostat -k sdb sdc sdd sde
Linux 2.6.38-999-generic (noisy)     01/31/2011     _x86_64_    (2 CPU)

avg-cpu:  %user   %nice %system %iowait  %steal   %idle
           7.51    0.00   11.86   18.09    0.00   62.55

Device:            tps    kB_read/s    kB_wrtn/s    kB_read    kB_wrtn
sdb             157.47       110.50      2448.63   40500708  897484716
sdd             149.80        42.17      2409.39   15456072  883103944
sdc             160.24        48.25      2526.69   17683264  926094568
sde             163.65        76.56      2439.40   28060388  894102580

root@noisy:~# iostat -k sdb sdc sdd sde
Linux 2.6.38-999-generic (noisy)     01/31/2011     _x86_64_    (2 CPU)

avg-cpu:  %user   %nice %system %iowait  %steal   %idle
           7.51    0.00   11.86   18.09    0.00   62.54

Device:            tps    kB_read/s    kB_wrtn/s    kB_read    kB_wrtn
sdb             157.52       110.48      2450.70   40503424  898471392
sdd             149.88        42.16      2411.66   15458072  884157636
sdc             160.30        48.24      2528.73   17685404  927080064
sde             163.73        76.55      2441.53   28063280  895109236

root@noisy:~#
```

The first iostat is of before the "dd", the second of after.

To my calculations a total of 3938MB was written while my replication level is at 3. In theory only 3*1024=3072MB should have been written. But there is a additional extra 28% of data which has been written to the disks.

This does not explain the very low performance of only 18MB/sec, but 28% of extra data being written seems a lot to me.

I benchmarked the individual OSD's and they are all around 55MB/sec, so that is fine.

Does Qemu-RBD do a lot of small files (with the extra 28% overhead) which brings the performance down and the I/O wait up?

#2 - 01/31/2011 11:22 AM - Wido den Hollander

- File bonnie.csv added

Oh, I forgot my bonnie++ results.

#3 - 02/01/2011 12:37 PM - Wido den Hollander

While running rsync I sometimes see these messages appearing inside the VM:

```
Feb  1 17:47:17 beta kernel: [ 7126.530741] flush-251:2: page allocation failure. order:0, mode:0x20
Feb  1 17:47:17 beta kernel: [ 7126.530750] Pid: 1086, comm: flush-251:2 Not tainted 2.6.32-28-server #55-Ubun
tu
Feb  1 17:47:17 beta kernel: [ 7126.530755] Call Trace:
Feb  1 17:47:17 beta kernel: [ 7126.530759]  <IRQ>  [<ffffffff810fa099>] __alloc_pages_slowpath+0x4a9/0x590
Feb  1 17:47:17 beta kernel: [ 7126.530780]  [<ffffffff810fa2f1>] __alloc_pages_nodemask+0x171/0x180
Feb  1 17:47:17 beta kernel: [ 7126.530789]  [<ffffffff8112d427>] alloc_pages_current+0x87/0xd0
Feb  1 17:47:17 beta kernel: [ 7126.530796]  [<ffffffff813dbab2>] try_fill_recv+0x182/0x200
Feb  1 17:47:17 beta kernel: [ 7126.530801]  [<ffffffff813dbcdd>] virtnet_poll+0x10d/0x160
Feb  1 17:47:17 beta kernel: [ 7126.530807]  [<ffffffff8147427f>] net_rx_action+0x10f/0x250
Feb  1 17:47:17 beta kernel: [ 7126.530814]  [<ffffffff8106d637>] __do_softirq+0xb7/0x1e0
Feb  1 17:47:17 beta kernel: [ 7126.530823]  [<ffffffff810936ba>] ? tick_program_event+0x2a/0x30
Feb  1 17:47:17 beta kernel: [ 7126.530829]  [<ffffffff810132ec>] call_softirq+0x1c/0x30
Feb  1 17:47:17 beta kernel: [ 7126.530833]  [<ffffffff81014cb5>] do_softirq+0x65/0xa0
Feb  1 17:47:17 beta kernel: [ 7126.530836]  [<ffffffff8106d4d5>] irq_exit+0x85/0x90
Feb  1 17:47:17 beta kernel: [ 7126.530844]  [<ffffffff815608f1>] smp_apic_timer_interrupt+0x71/0x9c
Feb  1 17:47:17 beta kernel: [ 7126.530848]  [<ffffffff81012cb3>] apic_timer_interrupt+0x13/0x20
Feb  1 17:47:17 beta kernel: [ 7126.530850]  <EOI>  [<ffffffff812a0d95>] ? __make_request+0x165/0x4a0
Feb  1 17:47:17 beta kernel: [ 7126.530861]  [<ffffffff8129f421>] ? generic_make_request+0x1b1/0x4f0
Feb  1 17:47:17 beta kernel: [ 7126.530865]  [<ffffffff810f5f05>] ? mempool_alloc_slab+0x15/0x20
Feb  1 17:47:17 beta kernel: [ 7126.530871]  [<ffffffff81436f11>] ? dm_merge_bvec+0xc1/0x140
Feb  1 17:47:17 beta kernel: [ 7126.530875]  [<ffffffff8129f7e0>] ? submit_bio+0x80/0x110
Feb  1 17:47:17 beta kernel: [ 7126.530882]  [<ffffffff81167382>] ? __mark_inode_dirty+0x42/0x1e0
Feb  1 17:47:17 beta kernel: [ 7126.530928]  [<ffffffffa00f6e74>] ? xfs_submit_ioend_bio+0x54/0x70 [xfs]
Feb  1 17:47:17 beta kernel: [ 7126.530945]  [<ffffffffa00f6f6b>] ? xfs_submit_ioend+0xdb/0xf0 [xfs]
Feb  1 17:47:17 beta kernel: [ 7126.530962]  [<ffffffffa00f7e2f>] ? xfs_page_state_convert+0x39f/0x720 [xfs]
Feb  1 17:47:17 beta kernel: [ 7126.530979]  [<ffffffffa00f830a>] ? xfs_vm_writepage+0x7a/0x130 [xfs]
Feb  1 17:47:17 beta kernel: [ 7126.530986]  [<ffffffff8110d265>] ? __dec_zone_page_state+0x35/0x40
Feb  1 17:47:17 beta kernel: [ 7126.530991]  [<ffffffff810fc0d7>] ? __writepage+0x17/0x40
Feb  1 17:47:17 beta kernel: [ 7126.530997]  [<ffffffff810fd257>] ? write_cache_pages+0x1d7/0x3e0
Feb  1 17:47:17 beta kernel: [ 7126.531001]  [<ffffffff810fc0c0>] ? __writepage+0x0/0x40
Feb  1 17:47:17 beta kernel: [ 7126.531006]  [<ffffffff810fd484>] ? generic_writepages+0x24/0x30
Feb  1 17:47:17 beta kernel: [ 7126.531023]  [<ffffffffa00f70fd>] ? xfs_vm_writepages+0x5d/0x80 [xfs]
Feb  1 17:47:17 beta kernel: [ 7126.531028]  [<ffffffff810fd4b1>] ? do_writepages+0x21/0x40
Feb  1 17:47:17 beta kernel: [ 7126.531032]  [<ffffffff81166676>] ? writeback_single_inode+0xf6/0x3d0
Feb  1 17:47:17 beta kernel: [ 7126.531037]  [<ffffffff81166da5>] ? writeback_sb_inodes+0x195/0x280
Feb  1 17:47:17 beta kernel: [ 7126.531044]  [<ffffffff81061671>] ? dequeue_entity+0x1a1/0x1e0
Feb  1 17:47:17 beta kernel: [ 7126.531048]  [<ffffffff811675c0>] ? writeback_inodes_wb+0xa0/0x1b0
Feb  1 17:47:17 beta kernel: [ 7126.531052]  [<ffffffff8116790b>] ? wb_writeback+0x23b/0x2a0
Feb  1 17:47:17 beta kernel: [ 7126.531057]  [<ffffffff81075fbc>] ? lock_timer_base+0x3c/0x70
Feb  1 17:47:17 beta kernel: [ 7126.531061]  [<ffffffff81076ad2>] ? del_timer_sync+0x22/0x30
Feb  1 17:47:17 beta kernel: [ 7126.531065]  [<ffffffff81167a19>] ? wb_do_writeback+0xa9/0x190
Feb  1 17:47:17 beta kernel: [ 7126.531068]  [<ffffffff810760d0>] ? process_timeout+0x0/0x10
Feb  1 17:47:17 beta kernel: [ 7126.531072]  [<ffffffff81167b53>] ? bdi_writeback_task+0x53/0xf0
Feb  1 17:47:17 beta kernel: [ 7126.531077]  [<ffffffff8110efa6>] ? bdi_start_fn+0x86/0x100
Feb  1 17:47:17 beta kernel: [ 7126.531081]  [<ffffffff8110ef20>] ? bdi_start_fn+0x0/0x100
Feb  1 17:47:17 beta kernel: [ 7126.531086]  [<ffffffff81084086>] ? kthread+0x96/0xa0
Feb  1 17:47:17 beta kernel: [ 7126.531090]  [<ffffffff810131ea>] ? child_rip+0xa/0x20
Feb  1 17:47:17 beta kernel: [ 7126.531094]  [<ffffffff81083ff0>] ? kthread+0x0/0xa0
Feb  1 17:47:17 beta kernel: [ 7126.531098]  [<ffffffff810131e0>] ? child_rip+0x0/0x20
Feb  1 17:47:17 beta kernel: [ 7126.531100] Mem-Info:
Feb  1 17:47:17 beta kernel: [ 7126.531104] Node 0 DMA per-cpu:
Feb  1 17:47:17 beta kernel: [ 7126.531108] CPU    0: hi:    0, btch:   1 usd:   0
Feb  1 17:47:17 beta kernel: [ 7126.531110] Node 0 DMA32 per-cpu:
Feb  1 17:47:17 beta kernel: [ 7126.531113] CPU    0: hi:  186, btch:  31 usd:  30
Feb  1 17:47:17 beta kernel: [ 7126.531120] active_anon:1457 inactive_anon:1616 isolated_anon:0
Feb  1 17:47:17 beta kernel: [ 7126.531122]  active_file:4171 inactive_file:229442 isolated_file:0
Feb  1 17:47:17 beta kernel: [ 7126.531123]  unevictable:0 dirty:9308 writeback:15301 unstable:0
Feb  1 17:47:17 beta kernel: [ 7126.531124]  free:1340 slab_reclaimable:8181 slab_unreclaimable:1908
Feb  1 17:47:17 beta kernel: [ 7126.531125]  mapped:1407 shmem:55 pagetables:330 bounce:0
Feb  1 17:47:17 beta kernel: [ 7126.531128] Node 0 DMA free:3996kB min:60kB low:72kB high:88kB active_anon:0kB
 inactive_anon:0kB active_file:0kB inactive_file:11548kB unevictable:0kB isolated(anon):0kB isolated(file):0kB
 present:15348kB mlocked:0kB dirty:0kB writeback:0kB mapped:0kB shmem:0kB slab_reclaimable:332kB slab_unreclai
mable:24kB kernel_stack:0kB pagetables:0kB unstable:0kB bounce:0kB writeback_tmp:0kB pages_scanned:0 all_unrec
```

```
laimable? no
Feb  1 17:47:17 beta kernel: [ 7126.531142] lowmem_reserve[]: 0 994 994 994
Feb  1 17:47:17 beta kernel: [ 7126.531147] Node 0 DMA32 free:1364kB min:4000kB low:5000kB high:6000kB active_
anon:5828kB inactive_anon:6464kB active_file:16684kB inactive_file:906220kB unevictable:0kB isolated(anon):0kB
 isolated(file):0kB present:1018060kB mlocked:0kB dirty:37232kB writeback:61204kB mapped:5628kB shmem:220kB sl
ab_reclaimable:32392kB slab_unreclaimable:7608kB kernel_stack:632kB pagetables:1320kB unstable:0kB bounce:0kB
 writeback_tmp:0kB pages_scanned:0 all_unreclaimable? no
Feb  1 17:47:17 beta kernel: [ 7126.531160] lowmem_reserve[]: 0 0 0 0
Feb  1 17:47:17 beta kernel: [ 7126.531165] Node 0 DMA: 1*4kB 1*8kB 3*16kB 1*32kB 5*64kB 10*128kB 1*256kB 0*51
2kB 0*1024kB 1*2048kB 0*4096kB = 3996kB
Feb  1 17:47:17 beta kernel: [ 7126.531177] Node 0 DMA32: 1*4kB 0*8kB 1*16kB 0*32kB 1*64kB 0*128kB 1*256kB 0*5
12kB 1*1024kB 0*2048kB 0*4096kB = 1364kB
Feb  1 17:47:17 beta kernel: [ 7126.531188] 233671 total pagecache pages
Feb  1 17:47:17 beta kernel: [ 7126.531191] 0 pages in swap cache
Feb  1 17:47:17 beta kernel: [ 7126.531193] Swap cache stats: add 0, delete 0, find 0/0
Feb  1 17:47:17 beta kernel: [ 7126.531196] Free swap  = 659448kB
Feb  1 17:47:17 beta kernel: [ 7126.531197] Total swap = 659448kB
Feb  1 17:47:17 beta kernel: [ 7126.537113] 262139 pages RAM
Feb  1 17:47:17 beta kernel: [ 7126.537118] 6224 pages reserved
Feb  1 17:47:17 beta kernel: [ 7126.537120] 235593 pages shared
Feb  1 17:47:17 beta kernel: [ 7126.537122] 22041 pages non-shared
Feb  1 17:47:17 beta kernel: [ 7126.537435] flush-251:2: page allocation failure. order:0, mode:0x4020
Feb  1 17:47:17 beta kernel: [ 7126.537441] Pid: 1086, comm: flush-251:2 Not tainted 2.6.32-28-server #55-Ubun
tu
Feb  1 17:47:17 beta kernel: [ 7126.537444] Call Trace:
Feb  1 17:47:17 beta kernel: [ 7126.537458]  [<ffffffff810fa099>] __alloc_pages_slowpath+0x4a9/0x590
Feb  1 17:47:17 beta kernel: [ 7126.537463]  [<ffffffff810fa2f1>] __alloc_pages_nodemask+0x171/0x180
Feb  1 17:47:17 beta kernel: [ 7126.537470]  [<ffffffff8112d427>] alloc_pages_current+0x87/0xd0
Feb  1 17:47:17 beta kernel: [ 7126.537475]  [<ffffffff81133467>] new_slab+0x2f7/0x310
Feb  1 17:47:17 beta kernel: [ 7126.537480]  [<ffffffff81135d11>] __slab_alloc+0x201/0x2d0
Feb  1 17:47:17 beta kernel: [ 7126.537488]  [<ffffffff8131d4b4>] ? vring_add_indirect+0x34/0x1c0
Feb  1 17:47:17 beta kernel: [ 7126.537492]  [<ffffffff811370ad>] __kmalloc+0x16d/0x1a0
Feb  1 17:47:17 beta kernel: [ 7126.537496]  [<ffffffff8131d4b4>] vring_add_indirect+0x34/0x1c0
Feb  1 17:47:17 beta kernel: [ 7126.537500]  [<ffffffff8131d868>] vring_add_buf+0x228/0x280
Feb  1 17:47:17 beta kernel: [ 7126.537506]  [<ffffffff81381509>] do_req+0x1a9/0x2c0
Feb  1 17:47:17 beta kernel: [ 7126.537510]  [<ffffffff8138165e>] do_virtblk_request+0x3e/0x90
Feb  1 17:47:17 beta kernel: [ 7126.537517]  [<ffffffff8129e9b3>] __generic_unplug_device+0x33/0x40
Feb  1 17:47:17 beta kernel: [ 7126.537521]  [<ffffffff81295e0a>] elv_insert+0x8a/0x200
Feb  1 17:47:17 beta kernel: [ 7126.537524]  [<ffffffff81295ff2>] __elv_add_request+0x72/0xd0
Feb  1 17:47:17 beta kernel: [ 7126.537529]  [<ffffffff812a0d5f>] __make_request+0x12f/0x4a0
Feb  1 17:47:17 beta kernel: [ 7126.537532]  [<ffffffff8129f421>] generic_make_request+0x1b1/0x4f0
Feb  1 17:47:17 beta kernel: [ 7126.537537]  [<ffffffff810f5f05>] ? mempool_alloc_slab+0x15/0x20
Feb  1 17:47:17 beta kernel: [ 7126.537542]  [<ffffffff81436f11>] ? dm_merge_bvec+0xc1/0x140
Feb  1 17:47:17 beta kernel: [ 7126.537546]  [<ffffffff8129f7e0>] submit_bio+0x80/0x110
Feb  1 17:47:17 beta kernel: [ 7126.537552]  [<ffffffff81167382>] ? __mark_inode_dirty+0x42/0x1e0
Feb  1 17:47:17 beta kernel: [ 7126.537578]  [<ffffffffa00f6e74>] xfs_submit_ioend_bio+0x54/0x70 [xfs]
Feb  1 17:47:17 beta kernel: [ 7126.537594]  [<ffffffffa00f6f6b>] xfs_submit_ioend+0xdb/0xf0 [xfs]
Feb  1 17:47:17 beta kernel: [ 7126.537609]  [<ffffffffa00f7e2f>] xfs_page_state_convert+0x39f/0x720 [xfs]
Feb  1 17:47:17 beta kernel: [ 7126.537625]  [<ffffffffa00f830a>] xfs_vm_writepage+0x7a/0x130 [xfs]
Feb  1 17:47:17 beta kernel: [ 7126.537631]  [<ffffffff8110d265>] ? __dec_zone_page_state+0x35/0x40
Feb  1 17:47:17 beta kernel: [ 7126.537636]  [<ffffffff810fc0d7>] __writepage+0x17/0x40
Feb  1 17:47:17 beta kernel: [ 7126.537640]  [<ffffffff810fd257>] write_cache_pages+0x1d7/0x3e0
Feb  1 17:47:17 beta kernel: [ 7126.537645]  [<ffffffff810fc0c0>] ? __writepage+0x0/0x40
Feb  1 17:47:17 beta kernel: [ 7126.537650]  [<ffffffff810fd484>] generic_writepages+0x24/0x30
Feb  1 17:47:17 beta kernel: [ 7126.537665]  [<ffffffffa00f70fd>] xfs_vm_writepages+0x5d/0x80 [xfs]
Feb  1 17:47:17 beta kernel: [ 7126.537670]  [<ffffffff810fd4b1>] do_writepages+0x21/0x40
Feb  1 17:47:17 beta kernel: [ 7126.537674]  [<ffffffff81166676>] writeback_single_inode+0xf6/0x3d0
Feb  1 17:47:17 beta kernel: [ 7126.537679]  [<ffffffff81166da5>] writeback_sb_inodes+0x195/0x280
Feb  1 17:47:17 beta kernel: [ 7126.537685]  [<ffffffff81061671>] ? dequeue_entity+0x1a1/0x1e0
Feb  1 17:47:17 beta kernel: [ 7126.537689]  [<ffffffff811675c0>] writeback_inodes_wb+0xa0/0x1b0
Feb  1 17:47:17 beta kernel: [ 7126.537692]  [<ffffffff8116790b>] wb_writeback+0x23b/0x2a0
Feb  1 17:47:17 beta kernel: [ 7126.537698]  [<ffffffff81075fbc>] ? lock_timer_base+0x3c/0x70
Feb  1 17:47:17 beta kernel: [ 7126.537702]  [<ffffffff81076ad2>] ? del_timer_sync+0x22/0x30
Feb  1 17:47:17 beta kernel: [ 7126.537706]  [<ffffffff81167a19>] wb_do_writeback+0xa9/0x190
Feb  1 17:47:17 beta kernel: [ 7126.537709]  [<ffffffff810760d0>] ? process_timeout+0x0/0x10
Feb  1 17:47:17 beta kernel: [ 7126.537713]  [<ffffffff81167b53>] bdi_writeback_task+0x53/0xf0
Feb  1 17:47:17 beta kernel: [ 7126.537718]  [<ffffffff8110efa6>] bdi_start_fn+0x86/0x100
Feb  1 17:47:17 beta kernel: [ 7126.537721]  [<ffffffff8110ef20>] ? bdi_start_fn+0x0/0x100
Feb  1 17:47:17 beta kernel: [ 7126.537726]  [<ffffffff81084086>] kthread+0x96/0xa0
Feb  1 17:47:17 beta kernel: [ 7126.537732]  [<ffffffff810131ea>] child_rip+0xa/0x20
Feb  1 17:47:17 beta kernel: [ 7126.537736]  [<ffffffff81083ff0>] ? kthread+0x0/0xa0
Feb  1 17:47:17 beta kernel: [ 7126.537739]  [<ffffffff810131e0>] ? child_rip+0x0/0x20
Feb  1 17:47:17 beta kernel: [ 7126.537741] Mem-Info:
Feb  1 17:47:17 beta kernel: [ 7126.537743] Node 0 DMA per-cpu:
Feb  1 17:47:17 beta kernel: [ 7126.537747] CPU    0: hi:    0, btch:   1 usd:   0
```

```
Feb  1 17:47:17 beta kernel: [ 7126.537749] Node 0 DMA32 per-cpu:
Feb  1 17:47:17 beta kernel: [ 7126.537752] CPU    0: hi:  186, btch:  31 usd:  30
Feb  1 17:47:17 beta kernel: [ 7126.537759] active_anon:1457 inactive_anon:1616 isolated_anon:0
Feb  1 17:47:17 beta kernel: [ 7126.537760]  active_file:4171 inactive_file:229442 isolated_file:0
Feb  1 17:47:17 beta kernel: [ 7126.537762]  unevictable:0 dirty:9308 writeback:15301 unstable:0
Feb  1 17:47:17 beta kernel: [ 7126.537763]  free:1340 slab_reclaimable:8181 slab_unreclaimable:1908
Feb  1 17:47:17 beta kernel: [ 7126.537764]  mapped:1407 shmem:55 pagetables:330 bounce:0
Feb  1 17:47:17 beta kernel: [ 7126.537767] Node 0 DMA free:3996kB min:60kB low:72kB high:88kB active_anon:0kB
 inactive_anon:0kB active_file:0kB inactive_file:11548kB unevictable:0kB isolated(anon):0kB isolated(file):0kB
 present:15348kB mlocked:0kB dirty:0kB writeback:0kB mapped:0kB shmem:0kB slab_reclaimable:332kB slab_unreclai
mable:24kB kernel_stack:0kB pagetables:0kB unstable:0kB bounce:0kB writeback_tmp:0kB pages_scanned:0 all_unrec
laimable? no
Feb  1 17:47:17 beta kernel: [ 7126.537780] lowmem_reserve[]: 0 994 994 994
Feb  1 17:47:17 beta kernel: [ 7126.537785] Node 0 DMA32 free:1364kB min:4000kB low:5000kB high:6000kB active_
anon:5828kB inactive_anon:6464kB active_file:16684kB inactive_file:906220kB unevictable:0kB isolated(anon):0kB
 isolated(file):0kB present:1018060kB mlocked:0kB dirty:37232kB writeback:61204kB mapped:5628kB shmem:220kB sl
ab_reclaimable:32392kB slab_unreclaimable:7608kB kernel_stack:632kB pagetables:1320kB unstable:0kB bounce:0kB
 writeback_tmp:0kB pages_scanned:0 all_unreclaimable? no
Feb  1 17:47:17 beta kernel: [ 7126.537799] lowmem_reserve[]: 0 0 0 0
Feb  1 17:47:17 beta kernel: [ 7126.537803] Node 0 DMA: 1*4kB 1*8kB 3*16kB 1*32kB 5*64kB 10*128kB 1*256kB 0*51
2kB 0*1024kB 1*2048kB 0*4096kB = 3996kB
Feb  1 17:47:17 beta kernel: [ 7126.537815] Node 0 DMA32: 1*4kB 0*8kB 1*16kB 0*32kB 1*64kB 0*128kB 1*256kB 0*5
12kB 1*1024kB 0*2048kB 0*4096kB = 1364kB
Feb  1 17:47:17 beta kernel: [ 7126.537827] 233671 total pagecache pages
Feb  1 17:47:17 beta kernel: [ 7126.537829] 0 pages in swap cache
Feb  1 17:47:17 beta kernel: [ 7126.537831] Swap cache stats: add 0, delete 0, find 0/0
Feb  1 17:47:17 beta kernel: [ 7126.537834] Free swap  = 659448kB
Feb  1 17:47:17 beta kernel: [ 7126.537835] Total swap = 659448kB
Feb  1 17:47:17 beta kernel: [ 7126.540126] 262139 pages RAM
Feb  1 17:47:17 beta kernel: [ 7126.540126] 6224 pages reserved
Feb  1 17:47:17 beta kernel: [ 7126.540126] 235593 pages shared
Feb  1 17:47:17 beta kernel: [ 7126.540126] 22041 pages non-shared
Feb  1 17:47:17 beta kernel: [ 7126.540126] SLUB: Unable to allocate memory on node -1 (gfp=0x20)
Feb  1 17:47:17 beta kernel: [ 7126.540126]   cache: kmalloc-2048, object size: 2048, buffer size: 2048, defau
lt order: 2, min order: 0
Feb  1 17:47:17 beta kernel: [ 7126.540126]   node 0: slabs: 31, objs: 248, free: 0
```

It seems that XFS can't flush enough at some point and runs out of allocated memory.

I could try and change "vm.min_free_kbytes", but the messages should not appear.

Could be related to the slow write performance I'm seeing, but I just wanted to report it.

#4 - 02/22/2011 06:09 AM - Wido den Hollander

I've spent a lot of time testing and finding out where this could come from, but it seems to be done now for no good reason.

I'm running 2.6.38-rc5 on all my OSD's (2 physical machines atm) with the btrfs code which is included in rc5. The ceph version I'm using is 4231cef68f534364767e98570e9b0785c4cc18e3 (Merge remote branch 'origin/max_commit_size').

Right now the write speeds seem OK (about 70MB/sec with replication = 3), but even better is de CPU usage. Where my machine "noisy" would have a load of about 7 or 8 during the rsync (See above), it's down to 2 ~ 3.

The virtual machine is also working much, much, much better. No more stalling in the VM, no messages in the dmesg either. rsync runs fine. Right now I've rsynced about 400GB of data multiple times and it is working fine.

Recovery actions also work much better, they run smooth and are not taking the whole system down, it actually keeps working during a recovery, a bit slower, but it works!

Took a lot of time, but this combination seems to be going fine.

#5 - 02/22/2011 08:34 AM - Sage Weil

- Status changed from New to Resolved

That's good news. I'm not really sure what was going wrong here before either. Let's see if this comes back.

Files

osd.0.log.gz	264 KB	01/28/2011	Wido den Hollander
bonnie.csv	2.79 KB	01/31/2011	Wido den Hollander