Subject: But why is the RAM gone?!
Posted by HubertD on Wed, 30 Jan 2008 11:39:14 GMT
View Forum Message <> Reply to Message

Hi,

our vz server seems to be suffering from some kind of memory leak since a kernel update from a homebrew 2.6.16-openvz to 2.6.18-ovz-stable in late november 2007.

I'm not completely sure that this is the kernel's or openvz's fault, but i also cannot find any process that allocates the missing amounts of memory.

I did put up a webpage containing "uname -a", "ps aux", "free -m", UBC values and some memory usages graphs over the last weeks:
http://www.denkmair.de/ramgone/

Please note that the mem usage drops every week are caused by reboots when/before the server starts swapping and becomes unresponsive. Reboots seem to be the only way to get my memory back.
I also tried to stop/restart only the containers and almost all running processes, but did not gain significant amounts of free memory.

Am I missing something?
Let me know if I can supply any additional information,

and thanks for your attention,

Hubert

[edit: added "lspci -v" and lsmod to webpage]

Subject: Re: But why is the RAM gone?!
Posted by maratrus on Wed, 30 Jan 2008 14:34:26 GMT
View Forum Message <> Reply to Message

Hi,

Have you any strange messages in logs or somewhere else?
I suppose that it is a common behavior. Kernel uses free memory for saving temporary objects (e.g. disk cache). So if your server will need memory kernel must give it to applications.

Subject: Re: But why is the RAM gone?!
Posted by HubertD on Wed, 30 Jan 2008 15:01:04 GMT
View Forum Message <> Reply to Message

Thank you for your answer!

I can't believe that it should be normal behaviour when used memory in "free" is considerably more (say, twice, thrice the size) than what rss size in "ps" sums up to. I don't speak about buffers and cache here, which of course take up most of the available memory.

But indeed, there are messages from the IBM RAID management program "ipssend" in dmesg/syslog:

```
sg_write: data in/out 4304/4304 bytes for SCSI command 0xd--guessing data in;
   program ipssend not setting count and/or reply_len properly
```

These messages have always been there, they appear every time the ipssend utitlity accesses the RAID hardware.
But together with the kernel update, I also installed nagios which calls the ipssend program regulary, 3 times every 5 minutes. Before, ipssend was only called manually a few times a day, might be possible that this creates the memory leak?
Silly me, didn't think about that.

I disabled the nagios check for now and I'll see if the leak's going to stop growing...

thank you very much for now,

Hubert

Subject: Re: But why is the RAM gone?!
Posted by HubertD on Wed, 30 Jan 2008 18:41:19 GMT
View Forum Message <> Reply to Message

I just called the ipssend command in question a few hundred thousand times and there seems to be no impact on the free memory.

So, I'm pretty sure that ipssend is not the origin of my memory leak.

Also, I wrote a small script to show the difference between ps and /proc/meminfo:

```
#!/usr/bin/perl -w
use strict;

my $pssum = 0;
open DATA, "/bin/ps -o rss= ax |";
while ( <DATA> ) { $pssum += $_; }
close DATA;
```

```
my $memtotal = 0;
my $memfree  = 0;
my $buffers  = 0;
my $cached   = 0;
open DATA, '</proc/meminfo';
while (<DATA>) {
        s/^\s+//;
        my @arr = split(/\s+/);
        if ($arr[0] eq 'MemTotal:')  { $memtotal = $arr[1]; }
        if ($arr[0] eq 'MemFree:')   { $memfree  = $arr[1]; }
        if ($arr[0] eq 'Buffers:')   { $buffers  = $arr[1]; }
        if ($arr[0] eq 'Cached:')    { $cached   = $arr[1]; }
}
close DATA;

my $used = $memtotal - $memfree - $buffers - $cached;
my $diff = $used - $pssum;
print "pssum: $pssum used: $used diff: $diff\n";
```

This shows a diff of ~1.7GB at the moment!

I have few knowledge of linux memory management, but this difference certainly doesn't seem normal to me?

Can somebody give me hints about how to find out more?

Hubert

Subject: Re: But why is the RAM gone?!
Posted by kir on Wed, 30 Jan 2008 19:39:07 GMT
View Forum Message <> Reply to Message

I run this script on my notebook (1G RAM, running Gentoo kernel at the moment, i.e. no VEs, no OpenVZ) and this is what it shows:

pssum: 445684 used: 321684 diff: -124000

I suspect that process RSS is not counting the memory used by kernel on behalf of the process, plus there's memory used by kernel itself. So sum of all RSS only gives you the lower estimation.

Such memory not accounted by RSS (neither 'cached' nor 'buffers' columns of /proc/meminfo) includes page tables, network buffers and other different kernel objects. I am not a kernel expert myself but will try to ask one tomorrow.

In any case, this can't work as an indication of a problem.

Subject: Re: But why is the RAM gone?!
Posted by HubertD on Thu, 31 Jan 2008 08:18:06 GMT
View Forum Message <> Reply to Message

Thank you Kir,

you're right.
I tried the script on several machines and always got a considerable negative or - at most - small positive diff reported.

However, I do not know another way to measure my problem, and I can't believe that a large positive value (>>1GB, as it was yesterday) should be okay.

I had to reboot the vz host this night (low memory would have killed it today otherwise), so it will take another week to fill up the RAM again.
I would be grateful for any hint on how to measure the gap more precisely or/and ideas about how to isolate the problem.

It's obvious that the openvz-stable kernel can't have such a big leak (>10MB/hour) in normal cases, but as I can't find a process allocating the missing RAM, it seems to be at least kernel-related to me.

I updated my website and added a "dmesg" print 4 hours after reboot of the vz host. Looks normal to me.

Thanks you for your time,

Hubert

Subject: Re: But why is the RAM gone?!
Posted by xemul on Fri, 01 Feb 2008 10:27:20 GMT
View Forum Message <> Reply to Message

Cached pages are included in the "busy" memory reported in /proc/meminfo, thus the "free" value is most often very low. However, beancounters do not account for such pages. The do not account for kernel memory allocated by drivers, disk request and some more as well, so if there's a leak in a driver we cannot definitely detect this.

The easiest way to find (ok - guess) which process consumes too much memory is to launch top and set the sorting by RSS value.
I'd pay more attention to 107 VE, since it shows to consume too many kmemsize and physpages, but if you launch top in ve0 it would be more than enough.

BTW, it would be nice if you also provide the kmemsize and physpapges graphs, not just privvmpages.

However, there can be some memory leak in your scsi (raid) driver. The best we can do now to

check this is to run these VEs on different hardware...

Thanks,
Pavel

Subject: Re: But why is the RAM gone?!
Posted by HubertD on Fri, 01 Feb 2008 12:03:37 GMT
View Forum Message <> Reply to Message

Thanks for your reply, Pavel,

I'll try to provide the kmemsize and physpages graphs, but since I didn't record that info in the first place, it's going to take some time...

I did of course search memory-consuming processes using "top", but cannot find any growing ones:
- After a reboot, there are >2GB "free" (-buffers,cache) memory
- The processes don't seem to grow over time.
  http://www.denkmair.de/ramgone/psmemgraph.png
  This is a graph that sums up "ps -o rss ax", and the total amount doesn't grow significantly.
- Only the "MemFree" info of /proc/meminfo is decreasing over time, while "Buffers" and "Cached" stay the same.
  Maybe I should keep track of other meminfo values, can somebody tell me which ones? ;-)

All together, I really suspect a kernel memory leak atm.
Unfortunately, simply changing the hardware node is not really a option, since this is a Co-Located server "in production" and it would be quite an effort to organize a equal replacement.

How difficult would it be to track down such a big (again, ~10MB/hour) kernel memory leak? Would this be possible on a production machine? (Nightly reboots are okay and so are kernel changes, as long as they don't have great impact on performance.)

Btw, this is all standard hardware, no funny exotic parts - an IBM eServer x345 with IBM/Adaptec ServeRaid controller, no other extensions.

Full lspci (and all other info) here:
http://www.denkmair.de/ramgone/

Thanks for your help, everybody!

Hubert

Subject: Re: But why is the RAM gone?!
Posted by xemul on Fri, 01 Feb 2008 12:50:44 GMT
View Forum Message <> Reply to Message

Quote:How difficult would it be to track down such a big (again, ~10MB/hour) kernel memory leak?

Well, a sequential dump of /proc/slabinfo file might help with it.

Upd: What is needed first of all is the number of object for each slab name (these names are in the /proc/slabinfo file's header), preferably sorted

This file contains statistics about what kind of internal kernel objects and how much of them are allocated. If we see some outrageous growth of some kind we'll be almost 100% sure that a leak takes place. Hopefully, we'll also know the origins of this leak

Thanks

Subject: Re: But why is the RAM gone?!
Posted by HubertD on Fri, 01 Feb 2008 13:03:44 GMT
View Forum Message <> Reply to Message

I've set up a cronjob that dumps slabinfo every 30minutes.
Let's see what happens in about 2 or 3 days?

Thanks so long!

Hubert

Subject: Re: But why is the RAM gone?!
Posted by HubertD on Mon, 04 Feb 2008 11:47:54 GMT
View Forum Message <> Reply to Message

Hi,

observing /proc/slabinfo for 4 days didn't take me any further

The memory leak keeps growing, but I can't find a process allocating that much RAM in "ps" and it also doesn't seem to be reflected in /proc/slabinfo in any way:

http://www.denkmair.de/ramgone/slabinfo.html

Free memory was
~2GB on 2008-02-01,
~1.5GB on 2008-02-02,
~1.3GB on 2008-02-03,
~1GB now (2008-02-04)

(I also attached a current dump of /proc/slabinfo at the end of

http://www.denkmair.de/ramgone/)

So, again, the server has ~1GB less free memory than 3 days ago, which is not used by buffers/cache, not used by any process (at least in terms of ps rss-size), and not reflected in the privvmpages values in UBC.

Where else could I search for it?

Or maybe I shouldn't care about RAM, better get some RUM and, when sober again, really invest into a new server?
Maybe with many more GBs of RAM, just to be on the safe side?

Greetings,

Hubert

Subject: Re: But why is the RAM gone?!
Posted by rickb on Mon, 04 Feb 2008 17:44:24 GMT
View Forum Message <> Reply to Message

Hi, I have been following your problem, seems quite strange and you have exhausted all means of tracking down memory usage (user and kernel). I would try to upgrade the kernel. I have excellent results with the 2.6.18 32bit EL5 PAE branch, using it on more then 100 servers without any memory leaks like this.

Rick

Subject: Re: But why is the RAM gone?!
Posted by xemul on Tue, 05 Feb 2008 11:46:07 GMT
View Forum Message <> Reply to Message

Hm... If kmemsize doesn't grow, then there's no memory leak.
This seems that the memory is mostly used for caching, but why are you sure that it is not?

Subject: Re: But why is the RAM gone?!
Posted by HubertD on Tue, 05 Feb 2008 15:52:50 GMT
View Forum Message <> Reply to Message

If the memory would be used for caching, then it should appear somewhere in "Buffers" or "Cached" in /proc/meminfo, I suppose?

The only thing that I'm really sure of is the following situation:

- After a reboot, free memory (including buffers and cache) reported by "free -m" starts at ~2GB
- This value is constantly decreasing over time, until there is almost none left after 6-7 days or so
- At this point, the server's load goes up dramatically (load>50), I suppose because of swapping, but I don't have prove for that.
- "Monit" on the server sees the high load and reboots the machine.

Before monit was installed, the machine became unresponsive and I wasn't able to log in via SSH or serial console any more.

I observed this cycle for about 10 times now.

During the cycle, the sum over "RSS size" of all processes (as returned by "ps aux") wouldn't increase, in fact, it decreases.

I'm almost at the end of this cycle (and my nerves ;) again and, did a snapshot of /proc/meminfo, /proc/user_beancounters, "ps -aux" etc. this noon.

I combined the data using openoffice calc, found out nothing new but at least it looks a little bit clearer as a spreadsheet:

(please excuse the picture, but you wouldn't like the super-messy html-export either ;)
The full document is available here: http://www.denkmair.de/ramgone/ubc.ods

If I understand the UBCs right, the sum(kmemsize+buffers+privvmpages) should reflect the total RAM usage on the machine, caches excluded.

So UBC tell me that I got ~2GB memory used.
"ps aux" sums up the process "resident sizes" to only about 1.4GB, which may go along with the 2GB from UBC.

But /proc/meminfo & "free" tell me that >3GB are actually used, buffers and caches not included.
THAT and the unresponsive machine when the free memory reported by /proc/meminfo reaches zero makes me sure that there REALLY IS a problem ;-)

Subject: Re: But why is the RAM gone?!
Posted by HubertD on Wed, 06 Feb 2008 11:58:08 GMT
View Forum Message <> Reply to Message

@xemul: did my last post convince you that there is some kind of leak? If not: Where do I go wrong?

Besides of changing the hardware node (which I do consider) and using other stable kernels

(which I definitly will give a try),
has anyone other ideas on how to track down the problem?

On kernels:
Would it be wise to test a precompiled enterprise-kernel
(linux-image-2.6.18-ovz-028stab051.1-enterprise debian package) on a 32bit xeon machine (4GB)?
Or should I better build it myself, e.g. something like the entnosplit/PAE version?
How stable is the 2.6.22-branch? Is it save enough for everyday use?

Subject: Re: But why is the RAM gone?!
Posted by xemul on Wed, 06 Feb 2008 13:25:37 GMT
View Forum Message <> Reply to Message

OK, I can imagine two more points of memory leaks.

The 1st is direct page allocations - they can be examined via /proc/buddyinfo file.
The 2nd one is vmalloc - get by cat /proc/meminfo | grep -i vmalloc command.

These are both kernel memory objects, but that would be strange if some driver uses them .

As the last attempt to understand what is going on is to wait till the free is close to zero, make the echo m > /proc/sysrq-trigger and then get the dmesg messages starting from 'SysRq: Show Memory' line up to the end - this will show us the overall memory state by the time of out-of-memory condition.

Upd: Can you please show us the config you compiled the kernel with.

Upd2:
Quote:Would it be wise to test a precompiled enterprise-kernel
(linux-image-2.6.18-ovz-028stab051.1-enterprise debian package) on a 32bit xeon machine (4GB)?
For 4Gb of ram regular kernel (smp/up) is more than enough.

Quote:How stable is the 2.6.22-branch? Is it save enough for everyday use?
2.6.22 was our development branch. It is no longer supported.

Subject: Re: But why is the RAM gone?!
Posted by HubertD on Wed, 06 Feb 2008 14:47:56 GMT
View Forum Message <> Reply to Message

Update: This is the current system status:

Sum(kmemsize,tcpsndbuf,tcprcvbuf,othersockbuf,dgramrcvbuf): 31,93 MB
privvmpages: 1919,61 MB

„ps aux“ RSS-Size: 1379,55 MB
„ps aux“ VSZ-Size: 4098,18 MB

MemTotal: 3861,42 MB
MemFree: 169,18 MB
Buffers: 66,18 MB
Cached: 249,43 MB

„free -m“ = MemFree+Buffers+Cached: 484,79 MB
MemTotal-MemFree-Buffers-Cached: 3376,63 MB

So, atm ~70MB more memory allocated in privvmpages (compared to yesterday), but 210MB less reported in "free -m".

buddyinfo is now:

```
Node 0, zone      DMA    335    139     66     12      0      0      0      1      1      1      0
Node 0, zone   Normal   8264  10177    806      9      1      1      1      0      1      1      0
Node 0, zone  HighMem   6955    744     25      1      0      1      1      0      0      0      0
```

cat meminfo | grep -i vmalloc

```
VmallocTotal:   114680 kB
VmallocUsed:      6944 kB
VmallocChunk:   107436 kB
```

Quote:2.6.22 was our development branch. It is no longer supported.
Sorry, I missed that. Also, I can't find a newer branch on the website. So should I stick with 2.6.18?

Subject: Re: But why is the RAM gone?!
Posted by HubertD on Wed, 06 Feb 2008 17:43:30 GMT
View Forum Message <> Reply to Message

I looked into the SysRq dmesg right now, a dump is available at:
http://www.denkmair.de/ramgone/index.html#sysrq

It doesn't really make me wiser, but maybe it helps somebody who does have kernel knowledge

What makes me wonder is the huge block (~2.7GB) of "inactive" pages. I am not a kernel hacker, and it's not easy for me to find out what is counted as "inactive".

For sure, this would be the file system cache (/proc/meminfo tells me ~230MB are "Cached" now),

and I also suppose the difference between privvmpages and oomguarpages (~1.2GB) are inactive pages.

What else?
Might the missing RAM make most of the other 1.3GB inactive pages?

greetings,

Hubert

Subject: Re: But why is the RAM gone?!
Posted by xemul on Thu, 07 Feb 2008 10:26:29 GMT
View Forum Message <> Reply to Message

Well. Large amount of inactive pages coupled with a rather big buffer_head kmemcache makes me think, that there are too many pages queued for disk output, but not complete yet.

This can be an IO problem somewhere at the iosched/driver/disk side.

Which iosched do you use? If CFQ, then you may try to switch to anticipatory one (called "as" in kernel) with this command:

echo "as" > /sys/block/<device-name>/queue/scheduler

so we will check whether this is an iosched or not (I believe it is not, but we should be sure)

As far as the driver is concerned - what kind of driver do you use - is this an out-of-tree driver, or the one provided with the kernel sources?

And I'm still worried with you driver warnings you mentioned

Subject: Re: But why is the RAM gone?!
Posted by HubertD on Thu, 07 Feb 2008 12:52:22 GMT
View Forum Message <> Reply to Message

All right, I'm going to use the anticipatory scheduler, just to be on the safe side.

But I don't think it could be a "normal" IO bottleneck as the server has low io-load and, after all, I'm graphing it

Here is a 24h io graph (data generated by "iostat"):
http://www.denkmair.de/ramgone/iostats/
the peaks show periodical rsync-backups of the live data on sda and sdb, the large reading phase on sdc at night is the daily offsite-backup.
Other than those, there is almost no ioload on the server, so waiting io requests could easily be

handled.

Concerning the raid driver...
cat /proc/scsi/ips/0

IBM ServeRAID General Information:

```
        Controller Type                 : ServeRAID 6M
        Memory region                    : 0xfb000000 (4096 bytes)
        Shared memory address            : 0xf8802000
        IRQ number                      : 16
        BIOS Version                    : 7.10.18
        Firmware Version                 : 7.10.18
        Boot Block Version               : 7.10.18
        Driver Version                  : 7.12.05
        Driver Build                    : 761
        Max Physical Devices             : 30
        Max Active Commands              : 64
        Current Queued Commands          : 0
        Current Active Commands          : 4
        Current Queued PT Commands       : 0
        Current Active PT Commands       : 0
```

This is the stock IBM ServeRAID driver from the kernel source, never used something else...
And with kernel 2.6.16, the machine has had an uptime of >400 days.
Also, the mentioned error messages have been there since day one and did not have any impact on stability.
They are gone now because I deactivated my raid status checks.

But, what's attracting my attention right now:
The displayed driver version differs from the firmware version.
No idea whether that is normal or could be a problem, I'm trying to find out...

Subject: Re: But why is the RAM gone?!
Posted by HubertD on Thu, 07 Feb 2008 22:58:42 GMT
View Forum Message <> Reply to Message

Nothing new from the ServeRAID frontier...
I have finally found the correct CD image @ibm.com, but upgrading a productive's system raid firmware at a remote site isn't exactly what I love doing

Anyways, swapping became unacceptable this evening and I decided to reboot the server again.
Here is the last dmesg listing before reboot:

SysRq: Show Memory
Mem-info:

DMA per-cpu:
cpu 0 hot: high 0, batch 1 used:0
cpu 0 cold: high 0, batch 1 used:0
cpu 1 hot: high 0, batch 1 used:0
cpu 1 cold: high 0, batch 1 used:0
DMA32 per-cpu: empty
Normal per-cpu:
cpu 0 hot: high 186, batch 31 used:137
cpu 0 cold: high 62, batch 15 used:52
cpu 1 hot: high 186, batch 31 used:14
cpu 1 cold: high 62, batch 15 used:54
HighMem per-cpu:
cpu 0 hot: high 186, batch 31 used:121
cpu 0 cold: high 62, batch 15 used:1
cpu 1 hot: high 186, batch 31 used:174
cpu 1 cold: high 62, batch 15 used:8
Free pages: 162456kB (44484kB HighMem)
Active:128331 inactive:791857 dirty:432 writeback:0 unstable:0 free:40676 slab:22831
mapped:10642 pagetables:1703
DMA free:7380kB min:68kB low:84kB high:100kB active:872kB inactive:716kB present:16384kB
pages_scanned:2455 all_unreclaimable? yes
lowmem_reserve[]: 0 0 880 4080
DMA32 free:0kB min:0kB low:0kB high:0kB active:0kB inactive:0kB present:0kB
pages_scanned:0 all_unreclaimable? no
lowmem_reserve[]: 0 0 880 4080
Normal free:110592kB min:3756kB low:4692kB high:5632kB active:50892kB inactive:607868kB
present:901120kB pages_scanned:0 all_unreclaimable? no
lowmem_reserve[]: 0 0 0 25600
HighMem free:31092kB min:512kB low:3928kB high:7344kB active:461588kB
inactive:2572840kB present:3276800kB pages_scanned:0 all_unreclaimable? no
lowmem_reserve[]: 0 0 0 0
DMA: 473*4kB 168*8kB 27*16kB 4*32kB 0*64kB 0*128kB 0*256kB 1*512kB 1*1024kB 1*2048kB
0*4096kB = 7380kB
DMA32: empty
Normal: 14858*4kB 5513*8kB 217*16kB 8*32kB 0*64kB 0*128kB 1*256kB 0*512kB 1*1024kB
1*2048kB 0*4096kB = 110592kB
HighMem: 6394*4kB 1666*8kB 73*16kB 0*32kB 0*64kB 1*128kB 1*256kB 0*512kB 0*1024kB
0*2048kB 0*4096kB = 40456kB
Swap cache: add 990247, delete 951427, find 3782192/3895280, race 0+816+43
Free swap = 1286308kB
Total swap = 1951856kB
Free swap: 1286308kB
1048576 pages of RAM
819200 pages of HIGHMEM
60052 reserved pages
148694 pages shared
38820 pages swap cached
432 pages dirty

0 pages writeback
10642 pages mapped
22831 pages slab
1703 pages pagetables
Top 10 caches:
buffer_head : size 45916160 objsize 52
size-4096(UBC) : size 2550080 objsize 4096
ext3_inode_cache : size 5619712 objsize 524
radix_tree_node : size 2785280 objsize 276
dentry_cache : size 4091904 objsize 144
filp : size 2088960 objsize 192
journal_head : size 1503232 objsize 52
vm_area_struct : size 3096576 objsize 84
size-2048 : size 1776320 objsize 2048
page_beancounter : size 10285056 objsize 32

And, just for the file, this is how /proc/meminfo looks after a reboot:

MemTotal: 3954096 kB
MemFree: 2531152 kB
Buffers: 74564 kB
Cached: 682844 kB
SwapCached: 0 kB
Active: 977500 kB
Inactive: 368632 kB
HighTotal: 3080044 kB
HighFree: 1795336 kB
LowTotal: 874052 kB
LowFree: 735816 kB
SwapTotal: 1951856 kB
SwapFree: 1951856 kB
Dirty: 2756 kB
Writeback: 0 kB
AnonPages: 581440 kB
Mapped: 126476 kB
Slab: 59988 kB
PageTables: 5708 kB
NFS_Unstable: 0 kB
Bounce: 0 kB
CommitLimit: 3928904 kB
Committed_AS: 2177148 kB
VmallocTotal: 114680 kB
VmallocUsed: 6944 kB
VmallocChunk: 107388 kB

Thanks for all your time,

Hubert

Subject: Re: But why is the RAM gone?!
Posted by den on Fri, 08 Feb 2008 13:04:52 GMT
View Forum Message <> Reply to Message

we have discussed the problem with Pavel and still not have good idea at all

First idea we have is to switch to RHEL5 based 028stab053.4, it can have a different set of drivers. Another idea is that we can remove some codepaths from question by installing Enterprise version of the kernel. This one will not have highmem on your host.

Regards,
Den