QGIS Application - Bug report #6448
Extremely slow shapefile reading over network

2012-10-03 03:23 AM - Regis Haubourg

Status:	Closed		
Priority:	Normal		
Assignee:			
Category:	Data Provider/OGR		
Affected QGIS version:	master	**Regression?:**	No
Operating System:		**Easy fix?:**	No
Pull Request or Patch supplied:	No	**Resolution:**	up/downstream
Crashes QGIS or corrupts data:	No	**Copied to github as #:**	15688

Description

Hi all,
QGIS really has a problem reading shapefile if this one is located over a network.
it's easy to reproduce:

Environment: QGIS commit:7ce99b8. Windows XP. LAN 100 Mb/s . HDD: SCSI 7400 Tr/s

File : multipolygon shp with mainly geometry (correctly indexed) . shp = 10 608 Ko, Dbf = 1100 Ko, shx = 22Ko

Results, Opening local shp on full extent:
- Raw file copy from Network to HDD : Less than 1s. network is not to blame
- ArcMAA 10: LOCAL: 3s   NETWORK : 5s
- QGIS 1.9 : LOCAL: 4s * NETWORK : 49 s!!*

I tested mapinfo tab format and qgis opens NETWORK in 8s, and LOCAL in 4s..

This is a blocker to me, already present before, **because I heard many disappointed users quitting QGIS**, "because it's too slow".. Indeed, this is often related to that issue.
Any idea? GDAL related ?
régis

History

#1 - 2012-10-03 05:32 PM - Even Rouault

Perhaps you can try this little experiment. Not a definitive fix, but would help diagnosing what's going on.

1) First make sure that the .shp, .shx and .dbf files are **read-only** (play with filesystem permissions) for the computer that reads the shapefile from the network
2) Define the VSI_CACHE environmement variable to YES
3) Possibly reboot your computer, and check in a "cmd" shell, that echo %VSI_CACHE% displays YES
4) Open QGIS and open the shapefile

#2 - 2012-10-04 02:30 AM - Regis Haubourg

Hi Even,
tests done and I found nothing different with or without VSI_CACHE set to YES. I use GDAL 1.9.1-2 with OSGEO4W. Anything else to test here?

#3 - 2012-10-04 02:43 AM - Even Rouault

Well, what you could try is to just time ogr2ogr in OSGeo4W Shell.

ogr2ogr target_local_file.shp source_local_file.shp

ogr2ogr target_local_file_2.shp x:\\yyyyy\\source_remove_file.shp

For the test with the remote file, check if the presence or absence of VSI_CACHE has some effect.

#4 - 2012-10-04 02:48 AM - Even Rouault

I meant x:\\yyyyy\\source_remote_file.shp (for the file in a network directory)

#5 - 2012-10-04 04:20 AM - Richard Duivenvoorde

we have been asked about this kind of issue also. I had a little chat on gdal-irc about it.

Not sure if it is usefull, but just want to add the irc log here:

see: http://irclogs.geoapt.com/gdal/%23gdal.2012-07-25.log

#6 - 2012-10-04 04:31 AM - Regis Haubourg

Further tests results:

1- I missed something in VSI_CACHE test. Setting files to Read-Only do improves speed: 1'10 without, 25s with. But it remains long.

2- ogr tests:
- ogr2ogr target_local_file.shp source_local_file.shp : < 1 s
- ogr2ogr target_local_file.shp source_local_file.shp :   9 s

Hope it helps

#7 - 2012-10-04 04:36 AM - Regis Haubourg

If needed, the polygon layer used for tests is available here : [http://dl.dropbox.com/u/72368800/data_sample_test.zip]

#8 - 2012-10-12 06:05 AM - hans windmuller

The problem lies in the shape files accompanied by large .dbf files.
I have tested 2 different shape files:
Shape file 1 consisting of 375555 features and a database sized 68 MB
Shape file 2 consisting of 132733 features and a database sized 162 MB

Both shape files were on a Windows 2003 server share at a distant office with a WAN SDSL connection of 10 Mbps.

Times for opening and rendering all features for Shape file 1:

QGIS 1.6   2’31”
QGIS 1.8   2’30”
ArcGIS 9.3.1   0’18”
ArcGIS 10.1   0’17”
DivaGIS   0’24”

Times for opening and rendering all features for Shape file 2:
QGIS 1.6   5’07”
QGIS 1.8   5’00”
ArcGIS 9.3.1   0’07”
ArcGIS 10.1   0’07”
DivaGIS   0’07”

In my opinion this shows that the speed for the Not Quantum GIS applications is related to the number of features whereas the speed of Quantum GIS is heavily related to the size of the .dbf file.

Richard Duivenvoorde suggested me to withdraw the .dbf file and then measure the speed. The not Quantum GIS applications are not able to do so, so I have only the times for the Quantum GIS applications:

Times for opening and rendering all features for Shape file 1:
QGIS 1.8   0’26”

Times for opening and rendering all features for Shape file 2:
QGIS 1.8   ’10”

Now Quantum GIS's speed is also related to the number of features. In my opinion Quantum GIS does not open the .dbf file on the remote storage (as the not Quantum GIS apps do) but downloads the .dbf file onto the client.

My question is if anyone of you can confirm this.

#9 - 2012-10-12 07:03 AM - Giovanni Manghi

hans windmuller wrote:

> *The problem lies in the shape files accompanied by large .dbf files.*
> *I have tested 2 different shape files:*
> *Shape file 1 consisting of 375555 features and a database sized 68 MB*
> *Shape file 2 consisting of 132733 features and a database sized 162 MB*
>
> *Both shape files were on a Windows 2003 server share at a distant office with a WAN SDSL connection of 10 Mbps.*
>
> *Times for opening and rendering all features for Shape file 1:*
> *QGIS 1.6   2’31”*
> *QGIS 1.8   2’30”*
> *ArcGIS 9.3.1   0’18”*
> *ArcGIS 10.1   0’17”*
> *DivaGIS   0’24”*
>
> *Times for opening and rendering all features for Shape file 2:*
> *QGIS 1.6   5’07”*
> *QGIS 1.8   5’00”*
> *ArcGIS 9.3.1   0’07”*

> *ArcGIS 10.1   0’07”*
> *DivaGIS   0’07”*
>
> *In my opinion this shows that the speed for the Not Quantum GIS applications is related to the number of features whereas the speed of Quantum GIS is heavily related to the size of the .dbf file.*
>
> *Richard Duivenvoorde suggested me to withdraw the .dbf file and then measure the speed. The not Quantum GIS applications are not able to do so, so I have only the times for the Quantum GIS applications:*
>
> *Times for opening and rendering all features for Shape file 1:*
> *QGIS 1.8   0’26”*
>
> *Times for opening and rendering all features for Shape file 2:*
> *QGIS 1.8   ’10”*
>
> *Now Quantum GIS's speed is also related to the number of features. In my opinion Quantum GIS does not open the .dbf file on the remote storage (as the not Quantum GIS apps do) but downloads the .dbf file onto the client.*
>
> *My question is if anyone of you can confirm this.*

there are differences also in a LAN?

#10 - 2012-10-12 07:56 AM - Regis Haubourg

Yes ! All my tests are made on Ethernet LAN.

#11 - 2012-10-14 11:51 PM - hans windmuller

And the same goes for me: there is even a difference between servers on the same location on the same (Ethernet) LAN.

#12 - 2012-10-17 05:04 AM - hans windmuller
- File QGIS_opening_shape_file_2.bmp added

In addition a screendump on which is shown that Quantum GIS generates networktraffic after first loading the geometry (.shp file) and later on (after clicking Zoom to Layer) loads the .dbf file and when finished renders the geometry.

#13 - 2012-11-27 07:38 AM - Regis Haubourg

More info:
- Files on a CIFS LAN (IBM storage M3600 powered by NetApp)
- LAN 100 MBit/s , but 1GBit/s on server side (network will be soon rewired)

- Win XP fresh Install of OSGEO4w qgis dev and qgis 1.8 DOES NOT REPRODUCE PROBLEM. 7s to open the same shp.
- 1.8 QGIS on another machine DOES REPRODUCE PROBLEM (40 s).
- same 1.8 QGIS installed on WIN 7 laptop DOES NOT Reproduce problem. Really fast I must say (4 s.)

This is strange.

#14 - 2012-11-30 02:49 AM - Regis Haubourg

Still trying to reproduce, I started installs of different packages on the same machine, with same profile and registry key.

Problem seems to come from configuration and not from library or qgis versions , since the package with the issue does not reproduce the issue after having reset profile, plugins and registry key.

#15 - 2012-11-30 07:16 AM - Regis Haubourg

Hi all,
I finally pointed two issues and need feedback to confirm. Strangely it does not seem to be related to network at all. Network delays effect were probably amplifying the issues

1- Suppressing following key speeds up shp drawing by two or more.
[HKEY_CURRENT_USER\\Software\\QuantumGIS\\QGIS\\Map]
"updateThreshold"=dword:00000005
It is the key storing snapping threshold + unit. What is triggered on layer loading? The default key I have stores no unit.

2- detection of shp projection often generates a custom CRS, since prj syntax is not recognised as proj4 syntax. When activating reprojection to the real projection (here it is 2154), proj 4 realize a reprojection to the same projection, which is slow.

I probably cumulated both issues + network perf issues.

Anybody confirm?
Thanks
Régis

#16 - 2012-12-30 09:51 AM - Giovanni Manghi

- Priority changed from High to Normal

#17 - 2013-04-04 09:47 AM - Radim Blazek

regis Haubourg wrote:

> *I finally pointed two issues and need feedback to confirm. Strangely it does not seem to be related to network at all. Network delays effect were probably amplifying the issues*

Good analysis.

> *1- Suppressing following key speeds up shp drawing by two or more.*
> *[HKEY_CURRENT_USER\\Software\\QuantumGIS\\QGIS\\Map]*
> *"updateThreshold"=dword:00000005*
> *It is the key storing snapping threshold + unit. What is triggered on layer loading? The default key I have stores no unit.*

The updateThreshold is number of features draw until screen is refreshed, it is set in "Options -> Rendering -> Number of features to draw before updating the display". If it is set to 0, display is not refreshed until all features are drawn. Display refresh is very slow, setting updateThreshold to a small number makes rendering very slow. For example, 10000 polygons 0.3s with updateThreshold=0 and 3.6s with updateThreshold=10.

We should restrict minimum updateThreshold to a bigger number, 1000 for example. Unfortunately 0 is used to disable display refresh so we cannot simply set min value for the control. We could add another independent checkbox but that is too invasive after feature freeze. My suggestion for 2.0 is to set spinbox step to 1000 and internally use 1000 as minimum (if user sets something between 0 and 1000).

Unfortunately I don't see any possible connection with slowdown (the amplification you suggested) when data are read over the network. Could you verify if changing updateThreshold from 0 to 1 really causes longer rendering time (the difference) over the network than for local files? In theory the difference between updateThreshold 0 and 1 for local and network should be the same (increased by constant number of screen updates).

> *2- detection of shp projection often generates a custom CRS, since prj syntax is not recognised as proj4 syntax. When activating reprojection to the real projection (here it is 2154), proj 4 realize a reprojection to the same projection, which is slow.*

Again, I don't see how it could be related to the network slowdown and why it did not affect MapInfo format.

> *I probably cumulated both issues + network perf issues.*

Please let us concentrate on the network issue first. That means disable other options which may influence rendering time: updateThreshold=0, no reprojection, single symbol renderer.

#18 - 2013-04-04 11:18 AM - Radim Blazek

I have done first simple test and I can just confirm that SHP over SMB is significantly slower than the same layer locally or Mapinfo TAB over SMB also with updateThreshold=0 and without reprojection. New issues should be created for those problems (updateThreshold and reprojection).

So we are at the beginning again, I think.

#19 - 2013-04-28 11:14 AM - Radim Blazek

I have done a lot of benchmarks and manual tests and here are my observations and conclusions:

- The problem is that OGR (if VSI cache is disabled) does not use any I/O buffer. If an operating system does not cache network files (Windows sometimes does not), it is obviously slow because files are read in small pieces, usually individual features.
- The most confusing and misleading is that network files on Windows are sometimes cached by the system and sometimes not. Files are cached if client gets opportunistic lock (OpLock). Once the lock is broken, files are not cached anymore. The rules when OpLock gets broken are not simple and sometimes are contra intuitive. In general, if a file is opened by a second process (may be running on the same or on another machine) it may or may not break the OpLock.
- VSI cache may act as a buffer, but it only supports reading from files opened in read only mode. QGIS opens all files in read write mode if possible, i.e. if files have read/write permission and update mode is supported by OGR provider (it is supported for Shapefile but not for Mapinfo TAB).
- VSI cache does not have any effect for Mapinfo TAB. Even Rouault pointed out that it is because Mapinfo driver uses the standard VSI API (which is just a thin wrapper around C standard FILE* API), whereas the Shapefile driver uses the VSI Virtual File API.
- DBF file is also read during rendering even if no attributes are used for rendering. I hope that this could be fixed in Shapefile provider because QGIS is giving to OGR list of attributes to be read. This is not core of the problem however.
- If update threshold (Options > Rendering > Number of features to draw before updating the display) is set to a small number, it can significantly slow down rendering.

Possible solutions

1. Immediate, partial.

- Set files permissions as read only on server
- Enable VSI_CACHE, in QGIS start up batch file (usually c:/osgeo4w/bin/qgis.bat or c:/osgeo4w/bin/qgis-dev.bat) set before qgis.exe is started environment variables, for example:

```
SET  VSI_CACHE=TRUE
SET  VSI_CACHE_SIZE=1000000
```

This is exactly what Even Rouault suggested in his very first reply and I believe that it should help for Shapefiles on XP and it was not proved because of various misleading circumstances (update threshold, OpLock broken/unbroken). It may be still slower (if VSI_CACHE_SIZE is smaller than file size), but it should not be more than about 50% slower, surely not many times slower.

2. QGIS workaround

We can enable VSI cache in start up batch by default, but that won't help if files are read/write. I am considering a possibility to open files always as read only and reopen as read/write only if necessary, but it is not that simple, it will be still slow when editing starts and reopening may cause other troubles.

3. OGR buffered I/O

OGR should use buffered I/O. It seems that VSI may work as buffer for reading. It has to be discussed with OGR developers if it is really suitable to be used as buffer and if it is possible to implement writing. Also possible increase of CHUNK_SIZE (32768) should be considered. This should be the true solution.

Tests

I would appreciate if you could test if my conclusions are correct. If you want to do tests, please follow this guidelines:

- Set "Options > Rendering > Number of features to draw before updating the display" to a number higher than number of features in the layer.
- Ensure that the files on server are not used during your tests by another process (regardless if it runs on the same machine or on another one). If you want to verify if OpLock gets broken by another process and if it has any effect on speed, just open the same file by another instance of QGIS.
- Let always load the layer first and then measure redraw time to avoid caching impact on first draw.
- Compare measured times with rendering time of the same layer rendered from local hard disk so that we can work with remote/local ratio.

Radim

#20 - 2013-04-28 11:29 AM - Even Rouault

I had a bit investigated about that issue some time ago and didn't report yet.

One thing is that when the OGR Shapefile driver will use both .shp and .dbf file, even when only drawing. When only drawing is needed, the OGR_L_SetIgnoredFields() API could be used to specify all the field names. In theory, one could think that this should be enough to avoid reading the DBF file. However, the Shapefile driver still needs reading the DBF file to check if the feature hasn't been deleted, because the flag to indicate if a feature is deleted is stored in the DBF file... The fact that other readers are faster would mean that they completely ignore the DBF file and potentially can draw deleted features ?

#21 - 2013-04-28 11:43 AM - Jürgen Fischer

rouault - wrote:

> *When only drawing is needed, the OGR_L_SetIgnoredFields() API could be used to specify all the field names.*

OGR_L_SetIgnoredFields() was introduced by Martin Dobias for that - so we already do that ;)
But most rendering also involves attributes for classification and labeling - so I suppose the case that only geometry is needed is quite rare.

#22 - 2013-04-28 11:59 AM - Radim Blazek

rouault - wrote:

> *The fact that other readers are faster would mean that they completely ignore the DBF file and potentially can draw deleted features ?*

Reading of DBF is not the core of the problem. It just multiplies the slowdown by two, because it reads each feature from both shp and dbf. Most probably it is not even important how many attributes are read (it may be however when they are processed in OGR and QGIS) because the slowdown comes from network latency not much from network speed.

#23 - 2013-04-29 12:56 AM - Radim Blazek

Discussion on GDAL devel list:
http://lists.osgeo.org/pipermail/gdal-dev/2013-April/036077.html

#24 - 2013-05-06 02:19 AM - Radim Blazek

I have added (commit:9222f152)

set VSI_CACHE=TRUE
set VSI_CACHE_SIZE=1000000

to qgis.bat (qgis.bat.tmpl). In case of performance problems with Shapefiles it should be sufficient to set files permissions as read only on server. Be aware however that changing of permissions to read only from read write for Mapinfo TAB may slow down rendering (OpLocks are broken for read only files, not for read write files if opened as read only, at least this was observed on Windows XP + Samba).

If you are interested in definitive solution of this problem, please sponsor development of write support in VSI cache of OGR (GDAL) library (www.gdal.org).

Thanks to Agence de l'eau Adour Garonne and Régis Haubourg who financially supported analysis of this problem.

#25 - 2013-07-02 12:41 PM - Radim Blazek

More comments in GDAL list:
http://lists.osgeo.org/pipermail/gdal-dev/2013-June/036545.html

#26 - 2014-06-21 03:53 AM - Jürgen Fischer

- Resolution set to up/downstream

- Status changed from Open to Closed

apparently an upstream issue now

#27 - 2014-06-21 03:53 AM - Jürgen Fischer

- Category set to Data Provider/OGR

Files

QGIS_opening_shape_file_2.bmp	2.26 MB	2012-10-17	hans windmuller