Subject: Incorrect SHA1 checksum for files 4GB+
Posted by Zbych on Mon, 21 May 2018 12:36:32 GMT
View Forum Message <> Reply to Message

Hi,

Did anyone test SHA1 using files 4GiB or larger?
When I calculate a sha1 for file 4294967295 bytes long, it is the same as returned by sha1sum from linux.
When file is 4294967296 bytes long (or longer) sha1 from Upp and sha1sum differs.

```
#include <Core/Core.h>

using namespace Upp;

CONSOLE_APP_MAIN
{
    constexpr int chunk = 1024 * 1024;
    constexpr int progress_interval = 1000;
//    static const char * filename = "/tmp/file4GB-.bin";
    static const char * filename = "/tmp/file4GB.bin";

    StdLogSetup(LOG_COUT | LOG_TIMESTAMP);
 RLOG("File '" << filename << "' sha1 calculation started");

    Sha1Stream sha1;
    FileIn file(filename);
    auto size = file.GetSize();
    auto last_progress = msecs();

    while (!file.IsError() && !file.IsEof()) {
        auto buff = file.Get(chunk);
        if (buff.GetCount() <= 0) break;
        sha1.Put(buff);
        if (msecs(last_progress) > progress_interval) {
            last_progress = msecs();
            if (size > 0) RLOG("Progress: " << 100 * file.GetPos() / size << "%");
        }
    }

    if (!file.IsError()) RLOG("SHA1: " << sha1.FinishString());
    else RLOG("File '" << filename << "' sha1 calculation interrupted by error: '" <<
file.GetErrorText() << "'");
}
```

Test files can be generated with following commands:

openssl rand 1073741824 > /tmp/file4GB-.bin
openssl rand 1073741824 >> /tmp/file4GB-.bin
openssl rand 1073741824 >> /tmp/file4GB-.bin
openssl rand 1073741823 >> /tmp/file4GB-.bin

openssl rand 1073741824 > /tmp/file4GB.bin
openssl rand 1073741824 >> /tmp/file4GB.bin
openssl rand 1073741824 >> /tmp/file4GB.bin
openssl rand 1073741824 >> /tmp/file4GB.bin

Subject: Re: Incorrect SHA1 checksum for files 4GB+
Posted by mirek on Sat, 16 Jun 2018 06:33:05 GMT
View Forum Message <> Reply to Message

This is really weird.

IMO, the only possible reason is a glitch in Stream::Get.

If you have a bit of time: Would it be possible to e.g. just read the file with Get(chucnk) and with some other methods (e.g. C++ streams or fopen) and compare?

Mirek

Subject: Re: Incorrect SHA1 checksum for files 4GB+
Posted by Zbych on Sat, 16 Jun 2018 09:19:29 GMT
View Forum Message <> Reply to Message

I've made some further tests:
1. reading and writing file using Upp's stream - no difference between original file and copy was detected.
2. replace Upp's sha1 with some random sha1 from git: https://github.com/clibs/sha1 - now checksum is the same as from sha1sum.
3. optimization level doesn't change result (both Upp's sha1 and git's)

Test app that calculates sha1 using Upp and git version at the same time is attached.

4GB file:

16.06.2018 11:11:27 File '/tmp/file4GB.bin' sha1 calculation started
16.06.2018 11:12:03 UPP SHA1: 1e4aca8da9f53733575cd351347b6a054a098bf5
16.06.2018 11:12:03 GIT SHA1: d43db14ac2c197da7d4d0048105f2af0bcb75f0b

$ sha1sum /tmp/file4GB.bin
d43db14ac2c197da7d4d0048105f2af0bcb75f0b /tmp/file4GB.bin

File smaller than 4GB:

16.06.2018 11:17:39 File '/tmp/file4GB-.bin' sha1 calculation started
16.06.2018 11:18:16 UPP SHA1: 30c411fa4ca4ed06b54ecb6143abcd39f2c32591
16.06.2018 11:18:16 GIT SHA1: 30c411fa4ca4ed06b54ecb6143abcd39f2c32591

$ sha1sum /tmp/file4GB-.bin
30c411fa4ca4ed06b54ecb6143abcd39f2c32591 /tmp/file4GB-.bin

File Attachments

1) sha1test.zip, downloaded 286 times

Subject: Re: Incorrect SHA1 checksum for files 4GB+
Posted by Zbych on Sat, 16 Jun 2018 09:44:43 GMT
View Forum Message <> Reply to Message

It appears that Upp is using exactly the same sha1 that I found on github (By Steve Reid). So the problem must be in Sha1Stream not sha1 itself.

Subject: Re: Incorrect SHA1 checksum for files 4GB+
Posted by mirek on Mon, 18 Jun 2018 10:20:40 GMT
View Forum Message <> Reply to Message

Should be now fixed. I have also added autotests for 5GB data with MD5, SHA1 and SHA256, all now seem to be OK.

Mirek

Subject: Re: Incorrect SHA1 checksum for files 4GB+
Posted by Zbych on Mon, 18 Jun 2018 19:28:39 GMT
View Forum Message <> Reply to Message

Thank you very much
