Computers & Geosciences 99 (2017) 155–158

Contents lists available at ScienceDirect

Computers & Geosciences

journal homepage: www.elsevier.com/locate/cageo

Research paper

Applying n-bit floating point numbers and integers, and the n-bit filter of HDF5 to reduce file sizes of remote sensing products in memory-sensitive environments

Stephan Zinke

EUMETSAT, Eumetsat Allee 1, 64295 Darmstadt, Germany

A R T I C L E I N F O

Keywords:
Data compression, Day and Night Band (DNB)
European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT)
Hierarchical Data Format version 5 (HDF5)
Suomi National Polar-orbiting Partnership (S-NPP)
Visible Infrared Imager Radiometer Suite (VIIRS)

A B S T R A C T

Memory sensitive applications for remote sensing data require memory-optimized data types in remote sensing products. Hierarchical Data Format version 5 (HDF5) offers user defined floating point numbers and integers and the n-bit filter to create data types optimized for memory consumption. The European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT) applies a compaction scheme to the disseminated products of the Day and Night Band (DNB) data of Suomi National Polar-orbiting Partnership (S-NPP) satellite's instrument Visible Infrared Imager Radiometer Suite (VIIRS) through the EUMETSAT Advanced Retransmission Service, converting the original 32 bits floating point numbers to user defined floating point numbers in combination with the n-bit filter for the radiance dataset of the product. The radiance dataset requires a floating point representation due to the high dynamic range of the DNB. A compression factor of 1.96 is reached by using an automatically determined exponent size and an 8 bits trailing significand and thus reducing the bandwidth requirements for dissemination. It is shown how the parameters needed for user defined floating point numbers are derived or determined automatically based on the data present in a product.

1. Introduction

In times of cheap memory (both RAM and disk) a tendency can be observed that remote sensing product files are not optimized for size. For applications, though, where file size is of importance, reduction mechanisms need to be exploited. Furthermore, access to smaller files is generally faster than to bigger files and can thus increase processing speed.

HDF5 (The HDF Group, 2011) is nowadays widely used for remote sensing products. While commonly internal compression via the HDF5 provided mechanisms "Shuffle" and "GZip" are applied, additional compression of resulting file sizes can be achieved by employing the standard, but not so well known, HDF5 features of user defined floating point numbers and integers, in conjunction with the application of the so called n-bit filter (The HDF Group, 2011).

At EUMETSAT, as part of the EARS services, remote sensing data from the VIIRS instrument onboard the NOAA/NASA polar orbiting satellite S-NPP are disseminated operationally over a broadband multicast system called EUMETCast. EUMETSAT has developed, in addition to the VIIRS Medium Resolution (M-Band) data service, a service to disseminate the DNB data of the VIIRS instrument.

Bandwidth for uploading of the data to the satellite broadcast dissemination system via DVB-S2 (EUMETCast) is costly, and it is desirable to reduce the original file sizes without loss of important information in the original data.

2. Method

HDF5 has a set of predefined data types, amongst others 32 bits IEEE Floating Point numbers and e.g. 16, 32 or 64 bits Integers. HDF5, however, also allows users to define floating point numbers and integers with bit-lengths which are not a multiple of 16. This feature allows customizing data types to the actual needs.

2.1. Floating point numbers representation

The IEEE 754 (IEEE, 2008) standard defines the encoding of floating point numbers, e.g. with 32 bits length as shown in Fig. 1. From least significant bit (LSB) to most significant bit (MSB) one can see 23 bits trailing significand (precision P=24 bits), 8 bits biased

Abbreviations: DNB, Day and Night Band; EARS, EUMETSAT Advanced Retransmission Service; EUMETSAT, European Organisation for the Exploitation of Meteorological Satellites; HDF5, Hierarchical Data Format version 5; NASA, National Aeronautics and Space Agency; NOAA, National Oceanographic and Atmospheric Administration; SDR, Sensor Data Records; S-NPP, Suomi National Polar-orbiting Partnership; VIIRS, Visible Infrared Imager Radiometer Suite

E-mail address: stephan.zinke@eumetsat.int.

http://dx.doi.org/10.1016/j.cageo.2016.11.003
Received 28 June 2016; Received in revised form 2 November 2016; Accepted 9 November 2016
Available online 13 November 2016

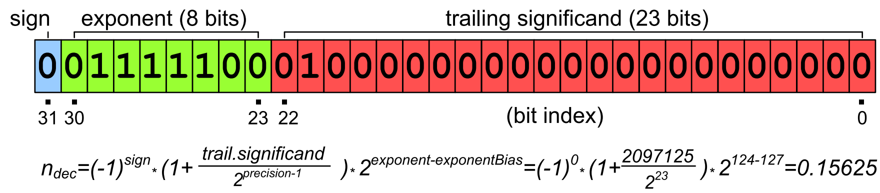

Fig. 1. Encoding layout of IEEE 754 32 bits floating point numbers.

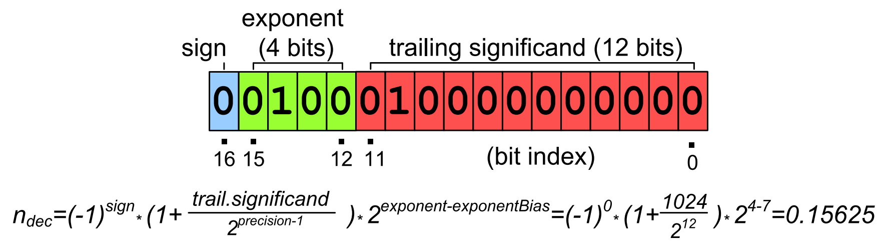

Fig. 2. Encoding layout of n-bit (17 bits) floating point numbers, example.

exponent and one sign bit; exponent bias is 127 (the exponent bias is explained in Section 2.1.2 below and is used to subtract the exponent number with the exponent bias). Note, that the trailing significand contains all the significand digits except the leading digit. The encoding of floating point numbers according to IEEE 754 implies the leading significand digit which is implicitly encoded in the biased exponent. Hence, the precision is one bit more than the total number of bits in the trailing significand (IEEE, 2008, p.9).

HDF5 allows additionally, following the standard encodings defined in IEEE 754, to define variable bit-length floating point numbers. Fig. 2 shows an example with 17 bits: 12 bits trailing significand (precision P=13 bits), 4 bits exponent and one sign bit; exponent bias is 7.

All parameters needed to define the n-bit floating point number in HDF5 need to be set by the defining user: sign-position, exponent-size, -position and -bias, trailing significand-size and -position. The following sections explain how the above mentioned parameters are determined or set.

2.1.1. Determining the needed exponent size automatically

The exponent size can be determined automatically and optimally based on the data represented in the data array.

Consider a system of normalized floating-point numbers F (B, P, L, U), where B is the base of the system, P is the precision of the system to P numbers, L is the smallest exponent representable in the system, and U is the largest exponent used in the system.

Then, there is a smallest positive normalized floating-point number, underflow level (Muller, 2010)

$$UFL = B^{L} \quad (1)$$

which has a 1 as the leading digit and 0 for the remaining digits of the significand, and the smallest possible value for the exponent.

There is a largest finite floating-point number, overflow level

$$OFL = (1 - B^{-P})(B^{U+1}) \quad (2)$$

which has $B - 1$ as the value for each digit of the significand and the largest possible value for the exponent.

Given a range of values $[|a|;|b|]$; a is the smallest number to be represented, b the largest; $B=2$ (binary representation).

L and U are to be determined based on given $[|a|;|b|]$, considering that practically L and U can only be integer values:

$$UFL = 2^{L} = a, \text{ and hence} \quad (3)$$

$$L = \lfloor \log_2 a \rfloor; \quad (4)$$

$$OFL = (1 - 2^{-P})*(2^{U+1}) = b, \text{ and hence} \quad (5)$$

$$U = \left\lceil \log_2\left(\frac{b}{1 - 2^{-P}}\right) - 1 \right\rceil. \quad (6)$$

The Range R of exponents needed is

$$R = U - L + 1. \quad (7)$$

Additionally, two special meaning exponents are required by IEEE 754 to represent infinities and Not-A-Number (NANs), and zero and subnormal numbers, i.e. one where all exponent bits are set and one where all exponent bits are unset. Thus, the real Range R' of exponents needed is

$$R' = U - L + 1 + 2. \quad (8)$$

The number of bits w needed to encode R' is therefore

$$w = \lceil \log_2 R' \rceil. \quad (9)$$

Example:

The range of data [1.0E-9; 1.6E-2] shall be represented optimally, P=6 (5 bits trailing significand). Using Eqs. (4), (6), (7), (8) and (9) we get $L = \lfloor -29.897 \rfloor = -30$ and $U = \lceil -6.943 \rceil = -6$; $R = 25$, $R' = 27$; $w = 5$.

Thus, the exponents needed, can be encoded with 5 bits.

2.1.2. Determining the needed exponent bias automatically

The exponent bias E_b is used to give the maximum possible range in the exponent without using a dedicated sign bit for the exponent, i.e. to make the biased exponent's range nonnegative. It is used to shift the exponent by the minimum exponent needed and as well accounting for the special meaning exponents required by IEEE 754:

$$E_b = -(L - 1). \quad (10)$$

In IEEE 754 the exponent bias for a 32 bits floating point number is 127.

The exponent bias used in the example above is 31.

2.1.3. Determining the needed trailing significand size

In contrast to the exponent size the trailing significand size cannot be automatically determined; the trailing significand size needs to be given based on the numerical accuracy required.

For example, the VIIRS DNB senses the data on-board with a resolution of 13 bits (low gain state), 13 bits (middle gain state), and 14 bits (high gain state) and specifies a calibration uncertainty between 5% (one half of maximum radiance for low gain state) and 100% (minimum radiance for high gain state) (NASA, 2014b). The measured Signal to Noise Ratio (SNR) performance of the DNB is given in (NASA, 2014a) and is 30–1000 over the specified measurement range ($4.0*10^{-9}$W/(cm^2*sr)–$3.0*10^{-2}$W/(cm^2*sr)).