Using SZIP and GZIP compression for HDF Datasets

July, 2005

1. Introduction

This note presents some examples of using SZIP [1] and GZIP [2] compression via HDF4 and HDF5 with real NASA datasets. Ths study used the *h4_compress_test.sh* and *h5_compress_test.sh* (see [5]) to compress example HDF-EOS (HDF4 and HDF5) data from NASA satellites. This script runs *hrepack* [3] (*h5repack* [4]) to copy all the objects in a file to a new file. In this project, *hrepack* (*h5repack*) was run repeatedly to apply GZIP [2] or SZIP [1] compression with all combinations of parameters that can be set for HDF4 (HDF5). The *hrepack* (*h5repack*) tool applied compression to every object that could be compressed, essentially arrays with numeric data that were larger than 1KB. The size of the output file was recorded for each setting. The time to run *hrepack* (*h5repack*) and the time to 'unpack' the compressed file were also recorded.

In all, there were 40 runs, 8 settings for GZIP, and 32 combinations of settings for SZIP. Table 1 shows the settings for GZIP, and Table 2 shows the settings for SZIP.

Two sets of data were used, MODIS L1B (HDF-EOS2/HDF4) and OMI L2 (HDF-EOS5/HDF5). One representative case of each is presented in this document.

Table 1. GZIP Compression Settings (see [2])

Parameter	Values	Description
Level	1-9	1= less compression, 9 = more compression

Table 2. SZIP Compression Settings (See [1,3,4]).

Parameter	Values	Description
Coding	Entropy Coding, Nearest Neighbor	See [1]
Pixels Per Block	Even number, from 2-32	Blocking factor

2. HDF4: MODIS Data

2.1. The Input Data: MODIS L1B data

This project used MODIS L1B data from 2004. These files all have the same structure, and the data values should be statistically similar across files. This paper presents the results for one file. We assume that other datasets from this data product would produce similar results.

This paper reports one file in detail,
MOD02QKM.A2004317.0000.004.2004317072459.hdf

The total size of the uncompressed file is 286,058,198 bytes. The file contains 16 objects, along with 50 global attributes, and 27 object attributes. Table 3 shows the objects and their sizes. There are 8 SDS arrays that can be compressed, for a total of 285,822,782 bytes (99.9%) that could be compressed. Note that the file contains 50 global attributes, amounting to at least 76,474 bytes, which cannot be compressed by the HDF4 library. Table 4 shows the global attributes.

2.3. Results

Figure 1 shows the size of the output file for the 40 conditions. The first 32 bars are SZIP, running from 2-32 Pixels per Block. The next 8 bars are GZIP, level 1-9. The rightmost bar is the size of the original uncompressed file, for reference.

For this dataset, GZIP achieved better compression than SZIP, with the expected slight improvement for higher values of the 'level' parameter.

For SZIP, it is clear that the 'Entropy coding' produced less compression than the 'Nearest Neighbor' for the same block size (at least for this dataset). Increasing the pixels per block increased the amount of compression. However, SZIP did not match or beat GZIP for the amount of compression.

Figure 2 shows the compression and decompression times. The first 32 bars are SZIP, and the next 8 are GZIP. The dark bars are the decompression time, and the light bars are the compression time. Overall, the decompression times were similar for all settings and for GZIP and SZIP. However, SZIP was uniformly much faster for compression than GZIP. 'Entropy Coding' gave slightly faster times than 'Nearest Neighbor' for equivalent block size.

2.4. Discussion

Overall, SZIP appears to run much faster than GZIP when compressing, but yields less compression for this sample dataset.

It should be noted that SZIP is designed to work best for floating point numbers. The majority of the data in the MODIS file is one or two byte integers.

Table 3. HDF4 objects in MOD02QKM.A2004317.0000.004.2004317072459.hdf (original size: 286,058,198).

Attr: global attributes, VG: VGroup, VD: Vdata table, SDS: Datasets, (*) means can be compressed (!) meas compressible but too small to bother with. VG, VD, and attributes cannot be compressed by HDF4.

Type	Object	Dimensions	Number type	Size (bytes)	Attrs
Attr	Global attributes (50)			76,474	
VG	MODIS_SWATH_Type_L1B				
VG	MODIS_SWATH_Type_L1B/Geolocation Fields				
SDS*	MODIS_SWATH_Type_L1B/Geolocation Fields/Latitude	2030x1345	Float 32	**10,921,400**	(3 attrs)
SDS *	MODIS_SWATH_Type_L1B/Geolocation Fields/Longitude	2030x1345	Float 32	**10,921,400**	"
VG	MODIS_SWATH_Type_L1B/Data Fields				
SDS *	MODIS_SWATH_Type_L1B/Data Fields/EV_250_RefSB	2 x 8120 x 5416	Uint 16	**175,911,680**	(14 attrs)
SDS *	MODIS_SWATH_Type_L1B/Data Fields/EV_250_RefSB_Uncert_Indexes	2x8120x5416	Uint 8	**87,955,840**	(7 attrs)
VD	MODIS_SWATH_Type_L1B/Data Fields/Band_250M	1		32	
VG	MODIS_SWATH_Type_L1B/Swath Attributes				
SDS*!	Noise in Thermal Detectors	16x10	Uint8	160	
SDS*!	Change in relative responses	16x10	Uint8	160	
SDS *	DC Restore Change for Thermal Bands	203x16x10	Uint8	**32,480**	
SDS *	DC Restore Change for Reflective 250m Bands	203x2x40	Int8	**16,240**	
SDS *	DC Restore Change for Reflective 500m Bands	203x5x20		**20,300**	
SDS *	DC Restore Change for Reflective 1km Bands	203x15x10		**30,450**	
VD	Level 1B Swath Metadata	203	64 bytes/ rec	12,992	
	Total Compressible			**285,822,782 (99.9%)**	

Table 4. Global Attributes in MOD02QKM.A2004317.0000.004.2004317072459.hdf (not compressible)

HDFEOSVersion	11	char
StructMetadata.0	32,000	char
CoreMetadata.0	16,515	char
ArchiveMetadata.0	2,921	char
Number of Scans	1	Int32
Number of Day mode scans	1	Int32
Number of Night mode scans	1	Int32
Incomplete Scans	1	Int32
Max Earth View Frames	1	Int32
%Valid EV Observations	38	Int32
%Saturated EV Observations	38	Float32
% L1A EV All Scan Data are Missing	1	Float32
% L1A EV RSB DN Not in Day Mode	490	Float32
% L1A EV DN Missing Within Scan	490	Float32
% Dead Detector EV Data	490	Float32
% Sector Rotation EV Data	490	Float32
% Saturated EV Data	490	Float32
% TEB EV Data With Moon in SVP	490	Float32
% EV Data Where Cannot Compute BG DN	490	Float32
% RSB EV Data With dn** Below Scale	490	Float32
% EV Data Where Nadir Door Closed	490	Float32
% EV Data Not Calibrated	490	Float32
Bit QA Flags Last Value	1	Uint32
Bit QA Flags Change	1	Uint32
Granule Average QA Values	50	Float32
Electronics Redundancy Vector	2	Uint32
Electronics Configuration Change	2	Uint32
Reflective LUT Serial Number and Date of Last Change	21	char
Emissive LUT Serial Number and Date of Last Change	21	char
QA LUT Serial Number and Date of Last Change	21	char
Focal Plane Set Point State	1	Int8
Doors and Screens Configuration	1	Int8
Reflective Bands With Bad Data	22	Int8
Emissive Bands With Bad Data	16	Int8
Noise in Black Body Thermistors	12	Int8
Noise in Average BB Temperature	1	Uint8
Noise in LWIR FPA Temperature	1	Uint8
Noise in MWIR FPA Temperature	1	Uint8
Noise in Scan Mirror Thermistor #1	1	Uint8
Noise in Scan Mirror Thermistor #2	1	Uint8
Noise in Scan Mirror Thermistor Average	1	Uint8
Noise in Instrument Temperature	1	Uint8
Noise in Cavity Temperature	1	Uint8
Noise in Temperature of NIR FPA	1	Uint8
Noise in Temperature of Vis FPA	1	Uint8
Dead Detector List	490	Int8
Noisy Detector List	490	Int8
Detector Quality Flag	490	Uint8
Earth-Sun Distance	490	Float32
Solar Irradiance on RSB Detectors over pi	330	Float32

Figure 1. Size of output file with different compression (MODIS HDF4)

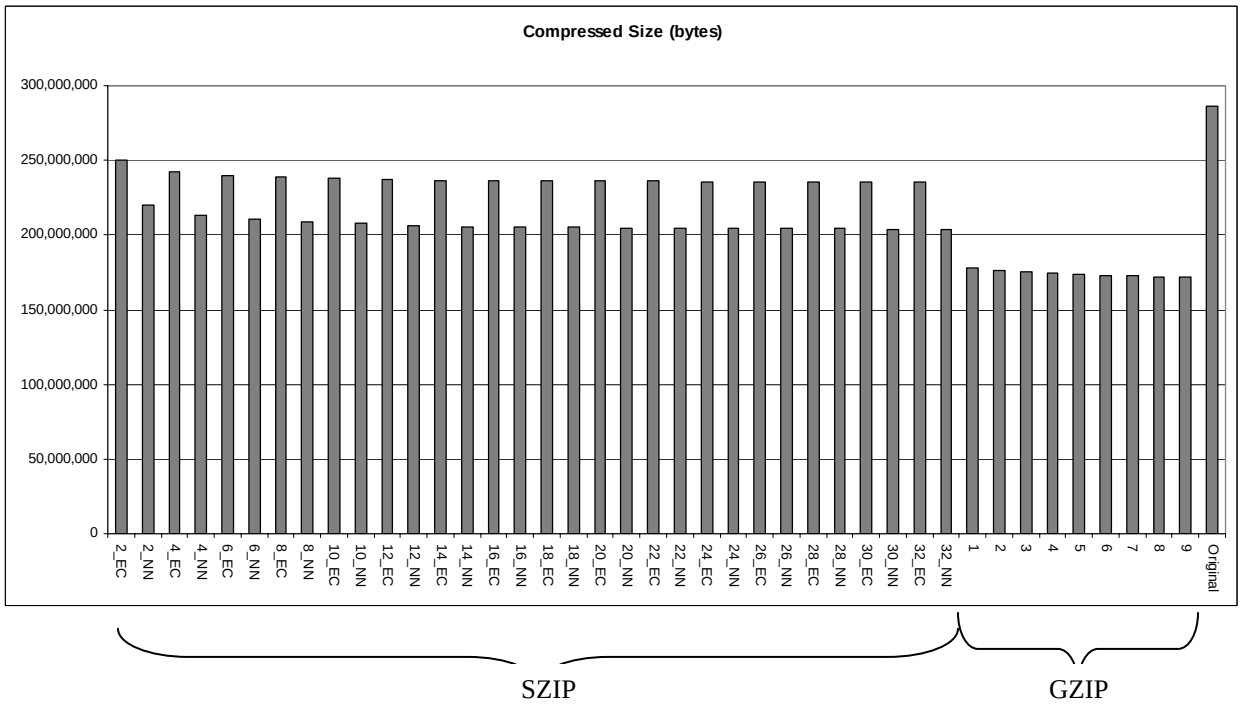
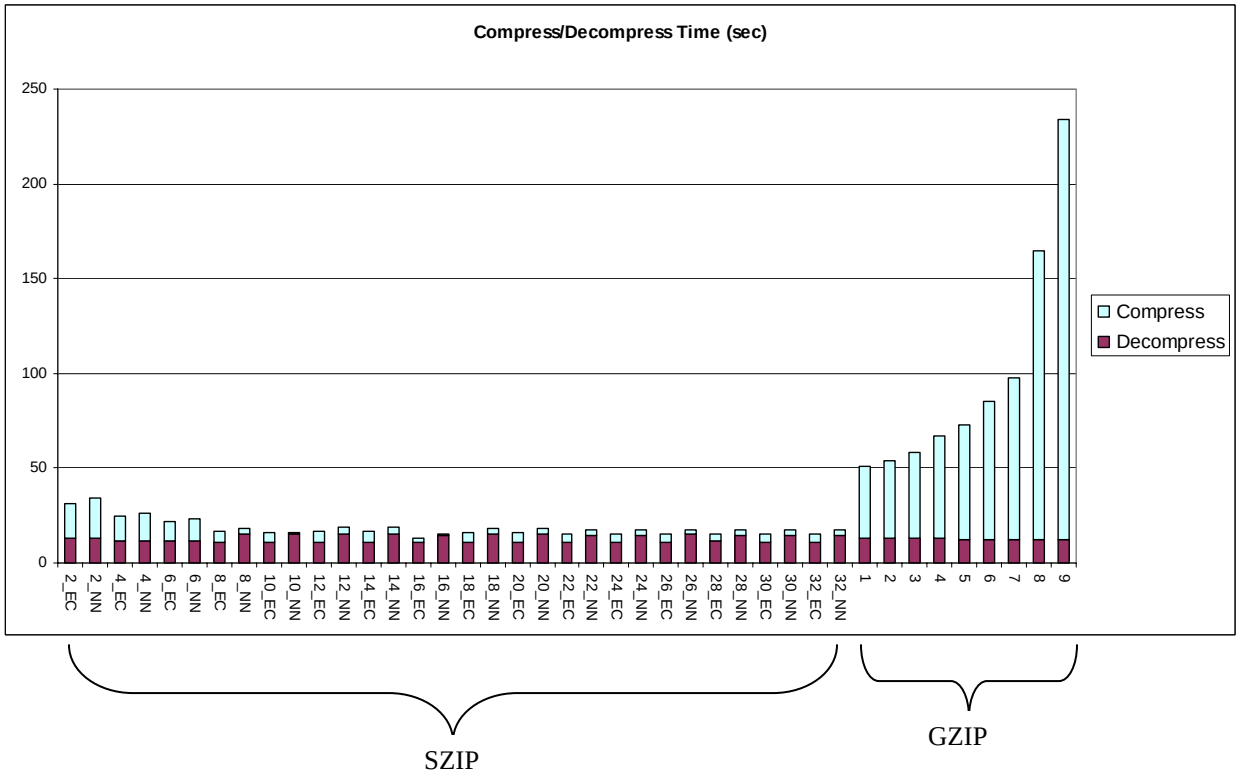


Figure 2. Time to Compress/Decompress the file (MODIS HDF4).



3. HDF5: OMI L2 Data

3.1. Input Data: OMI

This project used "OMI/Aura Ozone (O3) Total Column 1-Orbit L2 Swath 13x24km" data from 2005. These files all have the same structure, and the data values will be reasonably consistent. This paper presents the results for one file. We assume that other datasets from this data product would produce similar results.

This paper reports one file in detail,
OMI-Aura_L2-OMTO3_2005m0515t2259-o04437_v002-2005m0516t183605.he5

The total size of the uncompressed file is 40,383,606 bytes. The file contains 53 objects, along with 900 attributes.

Table 5 and Table 6 show the objects and their sizes. There are 41 Datasets that can be compressed, for a total of 40,166,104 bytes (99.4%) that could be compressed. Three datasets cannot be compressed, for a total of 163,110 bytes.

Each numeric dataset has six attributes (67 bytes per Dataset). Table 7 lists these attributes. There are 20 group attributes on several groups, for a total of 2,638 bytes. Table 8 lists the group attributes. These attributes are not compressed by HDF5.

3.2. Results

Figure 3 shows the size of the output file for the 40 conditions. The first 32 bars are SZIP, running from 2-32 Pixels per Block. The next 8 bars are GZIP, level 1-9. The rightmost bar is the size of the original uncompressed file, for reference.

For this dataset, GZIP achieved slightly better compression than SZIP, with the almost no improvement for higher values of the 'level' parameter.

For SZIP, it is 'Entropy coding' produced less compression than the 'Nearest Neighbor' for equivalent block size (for this dataset). Increasing the pixels per block increased the amount of compression. SZIP with entropy coding and 32 pixels per block was almost equivalent compression to GZIP.

Figure 4 shows the compression and decompression times. The first 32 bars are SZIP, and the next 8 are GZIP. The dark bars are the decompression time, and the light bars are the compression time. Overall, the decompression times were similar for all settings and for GZIP and SZIP. However, SZIP was uniformly much faster for compression than GZIP.

There was no clear pattern in the compression or decompression times for 'Entropy Coding' versus 'Nearest Neighbor' for equivalent block sizes.

Table 5. HDF5 objects in OMI-Aura_L2-OMTO3_2005m0515t2259-o04437_v002-2005m0516t183605.he5 (original size: 40,383,606). 1 of 2. (G: Group, D: Dataset, (*) means can be compressed (!) means compressible but too small to bother with. Groups, scalar datasets, attributes cannot be compressed by HDF5).

Type	Object	Elements	DT	Size
G	/			
G	/HDFEOS			
G	/HDFEOS/ADDITIONAL			
G	/HDFEOS/ADDITIONAL/FILE_ATTRIBUTES			
G	/HDFEOS/SWATHS			
G	/HDFEOS/SWATHS/OMI Column Amount O3			
G	/HDFEOS/SWATHS/OMI Column Amount O3/Data Fields			
D*	APrioriLayerO3	1496x60x11	F32	**3,949,440**
D*	AlgorithmFlags	1496x60	U8	**89,760**
D*	CloudFraction	1496x60	F32	**359,040**
D*	CloudTopPressure	1496x60	F32	**359,040**
D*	ColumnAmountO3	1496x60	F32	**359,040**
D*	InstrumentConfigurationId	1496	U8	**1,496**
D*	LayerEfficiency	1496x60x11	F32	**3,949,440**
D*	MeasurementQualityFlags	1496	U8	**1,496**
D*	NValue	1496x60x10	F32	**3,590,400**
D*	NumberSmallPixelColumns	1496	U8	**1,496**
D*	O3BelowCloud	1496x60	F32	**359,040**
D*	QualityFlags	1496x60	U16	**179,520**
D*	Reflectivity331	1496x60	F32	**359,040**
D*	Reflectivity360	1496x60	F32	**359,040**
D*	Residual	496x60x10	F32	**3,590,400**
D*	ResidualStep1	1496x60x10	F32	**3,590,400**
D*	ResidualStep2	1496x60x10	F32	**3,590,400**
D*	SO2index	1496x60	F32	**359,040**
D*	Sensitivity	1496x60x10	F32	**3,590,400**
D*	SmallPixelColumn	1496	I16	**2,992**
D*	StepOneO3	1496x60	F32	**359,040**
D*	StepTwoO3	1496x60	F32	**359,040**
D*	TerrainPressure	1496x60	F32	**359,040**
D*	UVAerosolIndex	1496x60	F32	**359,040**
D*!	Wavelength	10	F32	40
D*	dN_dR	1496x60x10	F32	**3,590,400**
D*	dN_dT	1496x60x10	F32	**3,590,400**

Table 6. HDF5 objects in OMI-Aura_L2-OMTO3_2005m0515t2259-o04437_v002-2005m0516t183605.he5 (original size: 40,383,606). 2 of 2. (G: Group, D: Dataset, (*) means can be compressed (!) means compressible but too small to bother with. Groups, scalar datasets, attributes cannot be compressed by HDF5).

Type	Object	Elements	DT	Size
G	/HDFEOS/SWATHS/OMI Column Amount O3/Geolocation Fields			
D*	GroundPixelQualityFlags	1496x60	U16	**179,520**
D*	Latitude	1496x60	F32	**359,040**
D*	Longitude	1496x60	F32	**359,040**
D*	RelativeAzimuthAngle	1496x60	F32	**359,040**
D*	SecondsInDay	1496	F32	**5,984**
D*	SolarAzimuthAngle	1496x60	F32	**359,040**
D*	SolarZenithAngle	1496x60	F32	**359,040**
D*	SpacecraftAltitude	1496	F32	**5,984**
D*	SpacecraftLatitude	1496	F32	**5,984**
D*	SpacecraftLongitude	1496	F32	**5,984**
D*	TerrainHeight	1496x60	I16	**179,520**
D*	Time	1496	F64	**11,968**
D*	ViewingAzimuthAngle	1496x60	F32	**359,040**
D*	ViewingZenithAngle	1496x60	F32	**359,040**
G	/HDFEOS/HDFEOS INFORMATION			
D*!	ArchivedMetadata	1	String 65,535	65,535
D*!	CoreMetadata	1	String 65,535	65,535
D* !	StructMetadata.0	1	string 32,000	32,000
	Compressible Data			40,166,104 (99.4%)

Table 7. Each numeric dataset has 6 attributes.

Attribute	Rank	DT	Size (bytes)
Units	1	String 2	2
Title	1	String 22	22
UniqueFieldDefinition	1	String 15	15
ScaleFactor	1	F64	8
Offset	1	F64	8
ValidRange	2	F32	8
MissingValue	1	F32	4

Table 8. Group attributes.

Attribute	Rank	DT	Size
GranuleDay	1	F32	4
GranuleMonth	1	F32	4
GranuleYear	1	F32	4
TAI93At0zOfGranule	1	F64	4
InputVersions	1	String 25	25
PGEVERSION	1	String 8	8
ProcessingCenter	1	String 8	8
InstrumentName	1	String 3	3
ProcessingHost	1	String 41	41
ProcessLevel	1	String 1	1
AuthorAffiliation	1	String 9	9
AuthorName	1	String 21	21
OrbitData	1	String 10	10
WavelengthOfAdjustment	10	F32	40
NVXAdjustment	600	F32	2,400
HDFEOSVersion	1	String 32	32
NumTimes	1	F32	4
NumTimesSmallPixel	1	F32	4
EarthSunDistance	1	F32	4
VerticalCoordinate	1	String 12	12

Figure 3. Size of output file with different compression (OMI HDF5).

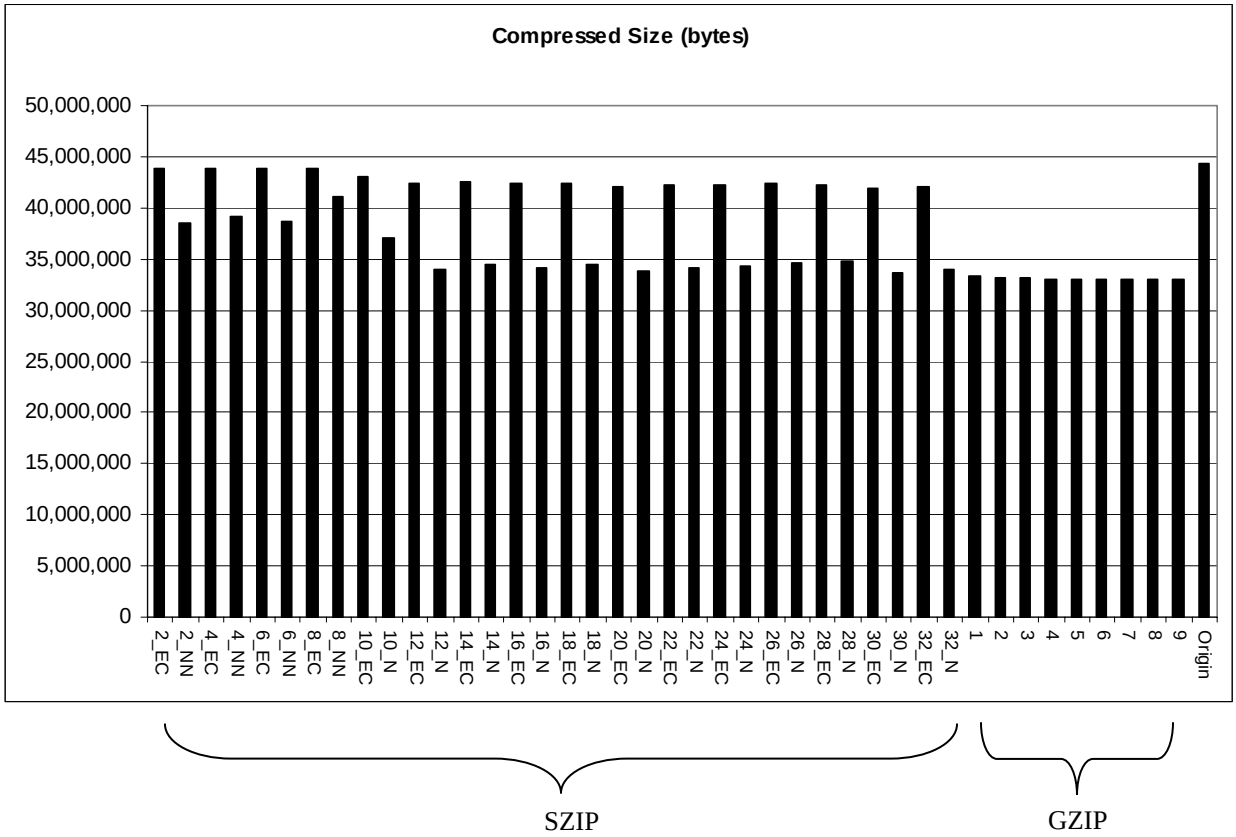
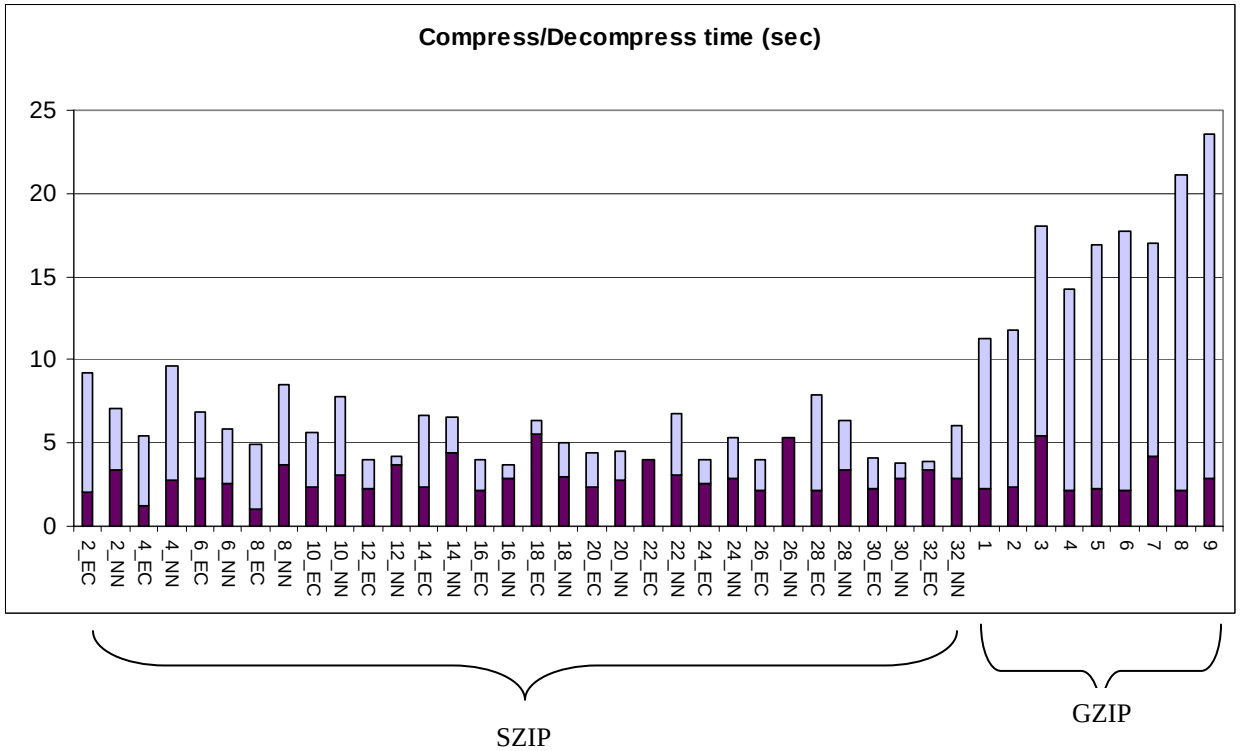


Figure 4. Time to Compress/Decompress the file (OMI HDF5).



3.3. Discussion

This example achieved relatively little compression with either method.

There are relatively many, relatively small datasets, no more than 4 MB in a single dataset. Since HDF5 compresses each dataset individually, the compression was applied to each of these small blocks, with corresponding overhead.

4. Summary

Overall, SZIP produced similar amounts of compression compared to GZIP, and was generally much faster in compressing.

The HDF4 example performed slightly better than the HDF5 example overall (e.g., when adjusted for the total size of the datasets). However, the HDF4 file contained one very large array, while the HDF5 example had no large arrays. This is likely the reason for the difference in compression size and speed.

It should be noted that these examples did not examine other parameters that might affect the compression, especially chunking. Also, it would be possible to write custom programs to compress the datasets individually, rather than using the repack utilities.

These examples illustrate the variability that may be encountered when compressing data with HDF. It can be very difficult to guess a priori what compression method or settings will work best for given data, or how well compression will work.

References

1. “Szip Compression in HDF Products”, http://hdf.ncsa.uiuc.edu/doc_resource/SZIP.
2. “Zlib”, http://www.zlib.net/
3. “NCSA HDF Tools”, http://hdf.ncsa.uiuc.edu/hdftools.html
4. “HDF5 Tools”, http://hdf.ncsa.uiuc.edu/HDF5/doc/Tools.html#Tools-Repack
5. “Compression Performance Evaluation Using Repack”