A Study of NetCDF as an Approach for High Performance Medical Image Storage

Marcone Magnus, Thiago Coelho Prado, Aldo von Wangenhein  
Post-graduate Program in Computer Science,  
Department of Informatics and Statistic,  
Federal University of Santa Catarina,  
Florianópolis, SC - Brasil  
E-mail: marcone, coeio, awangenh@inf.ufsc.br

Douglas D J de Macedo, M A R Dantas  
Post-graduate Program in Knowledge Engineering and Management,  
Department of Knowledge Engineering,  
Federal University of Santa Catarina,  
Florianópolis, SC - Brasil  
E-mail: macedo, mario@inf.ufsc.br

Abstract. The spread of telemedicine systems increases every day. The systems and PACS based on DICOM images has become common. This rise reflects the need to develop new storage systems, more efficient and with lower computational costs. With this in mind, this article discusses a study for application in NetCDF data format as the basic platform for storage of DICOM images. The study case comparison adopts an ordinary database, the HDF5 and the NetCDF to storage the medical images. Empirical results, using a real set of images, indicate that the time to retrieve images from the NetCDF for large scale images has a higher latency compared to the other two methods. In addition, the latency is proportional to the file size, which represents a drawback to a telemedicine system that is characterized by a large amount of large image files.

1. Introduction
The communication availability in recent years led the development and dissemination of applications aimed at meeting the needs of geographically dispersed users. One of these are the Telemedicine Systems, which have been applied as a way to provide access to the health of people or communities who are isolated or lack qualified medical treatment in one way or another. It’s use in these cases can reduce transportation costs and maximize the use of the technology park installed in hospitals and medical clinics [1].

In this way, the spread of PACS (Picture Archiving and Communication Systems) along with systems to support medical diagnosis based on images have been spreading steadily. This approach was applied to the project called RCTM Rede Catarinense de Telemedicina [2], [3]. This project was supported by the State of Santa Catarina Government and developed by the Telemedicine Laboratory [4], a member of the Cyclops Project [5] from Federal University of Santa Catarina, to support the lives of people around the state.
Currently, the RCTM is present in 286 cities of Santa Catarina, generating a monthly flow around 1.5 Terabytes of data, which are represented by more than 50000 medical exams performed by the population. The expectation for the coming years is to surpass the 150000 medical exams per month which will generate around 5 Terabytes per month of medical data. This reality leads to serious problems of data storage, that, for legal reasons, should be stored for a minimum period of 20 years.

Today the RCTM’s data is stored in a relational database, which has brought high storage costs with unsatisfactory performance. Thus, the development of new methods and techniques for storing data efficiently and with lower cost are being developed. This follows from the data replication to remote sites [6], [7], new data compression techniques and new ways of storing other data formats using [8].

In this article we will discuss a new format for storing data based on the data format NetCDF (Network Common Data Format) applied to telemedicine systems. The motivation for the development of this work is based on positive experience cited in the work developed by Macedo et. al. [8] that used data format HDF5 (Hierarchical Data Format 5) as the basic platform of data storage for the telemedicine system.

This article is structured in 6 sections, this being the first. Section 2 delineates a background of the research, seeking to base the fundamental concepts for understanding of the proposal. In this section are discussed telemedicine systems, DICOM Standard, HDF5 and NetCDF file formats. Section 3 exposes work-related research. Section 4 is the approach of the research. Section 5 exposes the experimental results, and the final section is the conclusion and future works.

2. Background

2.1. DICOM

In the mid 70’s, with the introduction of digital images from various modalities, manufacturers of medical equipment created their own standard of communication and data storage. As this type of digital image became popular and more companies manufactured equipment that generated it, the need for a single standard for all manufacturers increased. It was precisely why in the early 80’s that the ACR (American College of Radiology) together with the NEMA (National Electrical Manufacturers Association) created a single standard [9].

The first standard was published in 1985 and was called ACR-NEMA and was the first non-proprietary standard for storage and communication of medical data. In 1988 came the second version of the standard and in 1993 became what would be the third version that was called ACR-NEMA DICOM 3.0 [10]. Currently there are no plans for a new version but new features are added to the existing standard, the last being in 2011. DICOM, since then, is the basis for most of the PACS.

2.1.1. DICOM Information Object Definitions

The DICOM objects are generated by medical equipment and generally contain images and graphics. In addition the DICOM objects supports different types of data encapsulation such as waveforms, medical reports, PDFs, among others.

The components of the data structure of the DICOM described in the standard, that represents a study in the real world, are called IODs (Information Object Definitions). In the case of examinations which contains images such as CT scans and MRIs, there is information related to it, such as patient information, equipment and everything that is relevant for medical diagnosis. The IOD is structured in IE (Information Entities) that defines logical separations (the patient’s information, characteristics of the medical equipment, etc.), which in turn contains IOMs (Information Object Modules) that contain data elements. The data element is the unit of DICOM information and are containers that represent the data. There are elements for time
formats, date formats, UIDs, names, etc. In the standard are described for each IOD which IEs, IOMs and Data Elements are related [11].

2.1.2. DICOM Service Class  In addition to the data format described in the previous section, there are parts in the standard that formalize the communication protocol and service classes that can be supported. For this work we used the class provider of C-STORE service which is responsible for reception of medical imaging equipment for storage and interpretation [12].

2.2. Hierarchical Data Format
The HDF5 is the current version of a hierarchical data format created at NCSA (National Center for Supercomputing Applications). It is a self describing data format to store high amount of data. It can be employed in many areas like simulations involving physics, storage of geospatial data acquired by satellites or even in medicine. There are API for using the HDF5 data model in various languages like C, C++, Fortran and Java.

It resembles the XML format but the HDF5 stores data in binary format and lets you access data in specific parts of a file without having to load the entire content over random access. In addition, it supports aggregate data types that are compound atomic data type. It was designed to be able to execute partial I/O (read/write chunks) and parallel I/O (through MPI).

In contrast to relational database, hierarchical relashionship between data are expressed in a natural way. Relational database provides an excellent support for query based search on a comparison of field, but does not have good performance for sequential processing of all records or subsets of the data based coordinate-style lookup [13], like the extraction of a region of interest from an image.

2.3. NetCDF
Network Common Data Form (NetCDF) is a set of array-oriented interface for data access and a free collection of data access libraries for various languages like C, Fortran, C++, Java, among others. The NetCDF libraries support a machine-independent format for representing scientific data. Together, the interfaces and libraries support the creation, access and sharing of scientific data [14].

The NetCDF was developed and is maintained by Unidata Program from UCAR (University Corporation for Atmospheric Research). NetCDF is very popular data format for storing results of astronomical and physical climate simulations.
The NetCDF tool is older than HDF, the latter arose when in fact one of the goals was to replace the former in many applications. Because of this, many studies comparing these two tools were made. Many innovations that did not exist or did not work well in NetCDF were implemented in the HDF. In both latest versions of NetCDF (NetCDF4) the main change was the integration with HDF5. In conclusion, NetCDF4 uses many new features of HDF without losing compatibility with older versions.

The data model NetCDF4 that was implemented using a storage tier based on HDF5, works in all these limitations. In this model of enhanced data, a file has an initial group unnamed. Each group can contain one or more named variables, dimensions, attributes, groups and types. A variable is still a multidimensional array whose elements are all the same type, each variable may have attributes, and each format variable is specified by its size, which can be shared. However, in this enhanced model, one or more dimensions should be unlimited in size, so the data can be efficiently added to the variables along with any of these dimensions [14].

3. Related Works
The NetCDF has been a very popular and well established model for many years. There are several projects in various scientific fields that have based their systems or support for NetCDF. In this section we will see some of the projects and business entities that use or develop systems based on this tool.

The HALO (Halogen Occultation Experiment) was launched in atmospheric research satellite (SRVAS) in 1991, and, after a period of adjustment, it began scientific observations. The experiment uses solar occultation to measure vertical profiles of O3, HCl, HF, CH4, H2O, NO, NO2, and temperature versus pressure with an instantaneous vertical field of view of 1.6 km on Earth. The latitudinal coverage is 80 degrees S and 80 degrees N, over a year and includes extensive observations of the Antarctic region during spring.

The main objectives of this project are to study the dynamics of atmospheres of various, mainly polar, regions, study on the importance of the relationship between natural and anthropogenic sources of chlorine, measurement of atmospheric gases and aerosols and detailed analysis of development and recovery of the hole in the ozone layer in Antarctica. Data from the experiments of halo are organized in NetCDF format. The HALOE Data Viewer makes every data type with a menu-driven interface to assist in locating files based on date, time, type, and version of the data [15].

The LDEO (Lamont-Doherty Earth Observatory) is a research laboratory, at Columbia University, that complies with the geological and climatic behavior of the Earth. The observatory studies earthquakes, volcanoes and climate change in order to understand and predict the behavior of the Earth and how it affects and will affect society. Scientists observes Earth on a global scale, from its deepest interior to the outer layers of the atmosphere, on every continent and every ocean. They decipher the long history of the past, monitor the present and try to predict the future of Earth. The research institute converted all lasted geophysical data collected in 40 years by the scientists for the NetCDF data format [16].

The CHAMMP (Climate Change Prediction Program) is a program of the Department of Energy that conducts research in climate prediction. An important component of the program is that it unites CHAMMP emerging technologies in High Performance Computing for the development of computationally efficient models for precise numerical prediction of climate. Sponsored by the Institute of Biological and Environmental Research (OBER), the program involves an effort to develop computational methods capable of simulating future general circulation model of atmosphere and ocean. These programs are the core of advanced forecasting models that can be used to study climate change and they use NetCDF format as storage for their computer data [17].

The CSIRO (Commonwealth Scientific and Industrial Research Organisation) is a national
scientific agency in Australia and one of the largest and most diverse in the world. Among its areas of research are climate adaptation, preventive health, transformed energy, sustainable agriculture, and other ocean riches. The division of atmospheric research and the division for data coming from the ocean models uses for storage the NetCDF [18].

4. The Proposal
The main test to be performed with the proposed model of this work is a performance test comparing the results of the model that is used with HDF5. The chosen version was NetCDF4, which supports the hierarchical model of the HDF and fixes several bugs that the old versions of NetCDF had.

The DICOM data will be stored in a single file just like the HDF model does for hierarchical data representation. One can understand the model of this study through an analogy to the UNIX file system with its pattern of folders and subfolders. We define a folder structure based on the fixed table structure of the DICOM database currently used in RCTM. Each DICOM image will be inserted and organized according to this pre-defined structure.

This structure is divided into four levels. They were chosen to be implemented similarly to the way in [19], from where this model was chosen considering the information contained in images in DICOM format, as can be seen in Figure 2. The model enabled a satisfactory arrangement for the experiment, allowing a good view, an excellent performance and ease of access to data [19].

Originally the model had a hierarchical level that represented the 'Hospital' because it was considered a national scenario application, which does not correspond to reality today. In this work the level of the 'Hospital' was removed considering the fact that a patient can take exams in more than one health institution.

![Figure 2. Hierarchical data model used in NetCDF.](image)

As in the UNIX file system, the top of the hierarchy always starts with the “/” level and it is the only group in that level. At the second level we have the name of the patient. In this model there is still abstracted situations such as two patients with the same name, we know that this is possible and often but it is not the focus of this paper. The third level is the study and the fourth and final level is the series.

When we insert an image using DICOM, the application checks if there are already levels related to the image, if it does, the corresponding directory is open and have the data stored in an organized manner. If there are no directories, it is the first time that a patient is inserted
in an examination system, a new study or series, the respective directories to the new data elements are created.

5. Experimental Results
According to the chosen model, the medical images are stored in a single NetCDF file and within that file they are organized and stored in accordance with the determined hierarchy. It was noted that as the file sizes increased some operations were getting slower.

Figure 3 each bar represents the time spent for each image stored. Note that the last image took almost 30 times more than the first to be stored, although both have roughly the same size. This test was performed on a set of 100 images, all with approximately equal size.

![Figure 3. Behavior of image storage in NetCDF.](image)

Isolating each image store procedure, we notice that not all of them were directly related to the increasing time. We performed the same test but this time we measured the main methods runtime. The results show that although the opening operation time increased, the writing method presented a linear behavior, spending about the same time running on all images.

6. Conclusions and Future Works
The results from the test showed a behavior that was not expected in this work. The model implemented is similar to the one in [19], both data formats share a common layer that handles the hierarchical portion of the tool. Despite the HDF5 promising behavior, the NetCDF does not meet the expectations, since the storage time increases as the file grows.

Using a single file to store all the information is not a viable alternative to the problem faced in the RCTM. Though this means that the tools is not suitable for working with this model, it does not mean that it is not suitable for working in this environment.

We contacted the NetCDF development team and we were informed that in order to accelerate access to data recorded the entire NetCDF file is loaded into memory, in HDF5 this does not happen, it only loads the requested data. Thus we assured the unsuitable use of this tool for this model.

A new model is being studied to replace the hierarchical model storage of medical images with NetCDF. This new model should consider the tool limitations detected in this work, as well the advantages of the old format that is not hierarchical based.
7. References

[1] McNeill K M, Weinstein R S and Holcomb M J 2002 *Journal of American Medical Informatics Association* URL: http://www.pubmedcentral.gov/articlerender.fcgi?artid=61325

[2] Maia R, von Wangenheim A and Nobre L 2006 *19th IEEE International Symposium on Computer-Based Medical Systems, 2006. CBMS 2006.* pp 495–500

[3] Wallauer J, Macedo D D J, Andrade R and von Wangenheim A 2008 *IT Professional* 10 12–17 ISSN 1520-9202

[4] Laboratório de telemedicina Available in: http://www.telemedicina.ufsc.br. Accessed in: 07/14/2011. URL http://www.telemedicina.ufsc.br

[5] The cyclops project Available in: http://www.cyclops.ufsc.br. Accessed in: 07/14/2011 URL http://www.cyclops.ufsc.br

[6] Macedo D D J, Perantunes H W G, Andrade R, von Wangenheim A and Dantas M A R 2008 *21th IEEE International Symposium on Computer-Based Medical Systems, 2008. CBMS 2008.* pp 495–500

[7] Macedo D D J, Perantunes H W G, Comunello E, von Wangenheim A and Dantas M A R 2008 *IEEE Symposium on Computers and Communications – ISCC, 2008*

[8] Macedo D, Von Wangenheim A, Dantas M and Perantunes H 2009 *International Journal of High Performance Systems Architecture* 2 99–106

[9] NEMA 2009 Digital imaging and communications in medicine (dicom) part 1: Introduction and overview Tech. rep. National Electrical Manufactures Association.

[10] Mildenberger P, Eichelberg M and Martin E 2002 *European radiology* 12 920

[11] NEMA 2009 Digital imaging and communications in medicine (dicom) part 3: Information object definitions Tech. rep. National Electrical Manufactures Association.

[12] NEMA 2009 Digital imaging and communications in medicine (dicom) part 4: Service class specifications Tech. rep. National Electrical Manufactures Association.

[13] group H 2010 Hierarchical data format Acessado em: 28 abr. 2010 URL: http://www.hdfgroup.org/

[14] UNIDATA 2010 Network common data format Acessed in: 28 abr. 2010 URL http://www.unidata.ucar.edu/software/netcdf/

[15] NASA 2010 Haloe Acessed in: 14/07/2011 URL http://haloe.gats-inc.com/about/index.php

[16] LDEO 2010 Lamont-doherty earth observatory Acessed in: 14/07/2011 URL http://www.ldeo.columbia.edu/

[17] CHAMMP 2010 Climate change prediction program Acessed in: 14/07/2011 URL http://www.csm.ornl.gov/champp/champp.html

[18] CSIRO 2010 Commonwealth scientific and industrial research organisation Acessed in: 14/07/2011 URL http://www.its.csiro.au/csiro/channel/

[19] de Macedo D, Comunello E, von Wangenheim A and Dantas M 2008 *Proceedings of the 14th Brazilian Symposium on Multimedia and the Web (ACM)* {pp} 20–27