CSAR: A new way forward with high volume bathymetry

Karen Hart, CARIS, USA
Karen Cove, CARIS, Canada
Corey Collins, CARIS, Canada

Introduction

Efforts to decrease the time and cost associated with completing the ping to chart process are in direct conflict with the growing trend of having to work with very large surveys in both data volume and area coverage. This poses unique processing, storage, and distribution challenges.

Bathymetric processing software is evolving to accommodate the challenge of the increased data volumes and provide new modalities for interacting with it. In an effort to address these trends, improve the productivity of its users, and permit more sophisticated data analysis, CARIS® has developed a new storage technology that will allow its users to store, process, and visualize very large volumes of data in a unique approach.

Bathymetric data is stored as either high-density gridded datasets or lower density vector point datasets, both of which have very high storage space requirements. Although data collected in most modern surveys ends up as gridded datasets, the vast majority of existing or historic datasets can still be considered sparse data that is preferably stored as points. A bathymetry data management system has to accommodate both types of data and give the user opportunity to use both together to produce optimum coverage and quality for products.

CARIS has traditionally managed both gridded and point data, but it has become apparent that the formats used were not well suited to the task of handling the much larger volumes of data collected by modern sensors, such as multibeam and interferometric sonars, and LiDAR sensors. This fact, coupled with a demand for decreased ping-to-product times has prompted CARIS to re-examine traditional approaches for data storage.

New Approach to handling High Volume Bathymetry

Starting with a clean sheet design, CARIS is implementing the next generation storage technology with high volume data and fast visualization in mind. The new storage format will facilitate efficient spatial queries and data access over a network, and is designed to allow users to open a file and see their data as quickly as possible. This new technology will be incorporated into CARIS applications such as HIPS and SIPS and Bathy DataBASE. CARIS is also extending its web-based applications to take advantage of these visualization capabilities, making it much easier to access and view high volume bathymetry over the web.

[1] CARIS USA, Alexandria, VA USA
[2] CARIS Canada, Fredericton, NB CANADA

The CSAR Framework

CARIS designed the CSAR Framework to handle large volumes of multi-dimensional data by partitioning into pieces called “chunks”, each of which is given a unique key that can be used to retrieve it from a storage device such as a file or a database. A collection of chunks is called a Set. Because I/O system performance generally varies with the number of reads or writes required to move data to and from storage, the amount of data stored in each chunk will generally dictate the system’s performance. A diagram of the application technology stack is shown in Figure 1.

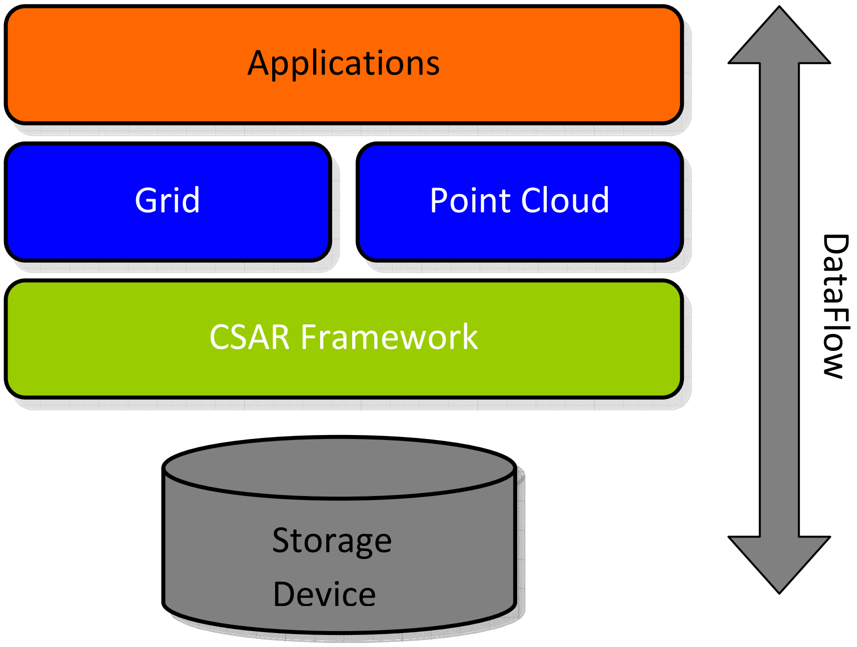

Figure 1: CARIS Application Technology Stack

CARIS has re-implemented its data structures for gridded and point data using the CSAR framework to accommodate the very large volumes of data now produced by the bathymetric workflow. The data structures are described in detail below.

Data Structures

In addition to partitioning all datasets into manageable chunks, both gridded and point data structures have been made “multi-resolution.” This means the data is stored at several resolutions simultaneously to facilitate very rapid access. The original data represents the highest available resolution, and lower resolution approximations are automatically constructed as data is added.

An overview of an entire dataset can be retrieved by loading only the coarsest level data from a storage device. Small subsets of high resolution data can be loaded just as quickly. An example of a partitioned, multi-resolution gridded dataset is shown below. The original grid was constructed with 10 meter spacing; the 20m and 40m grids were constructed as approximations of the original grid.

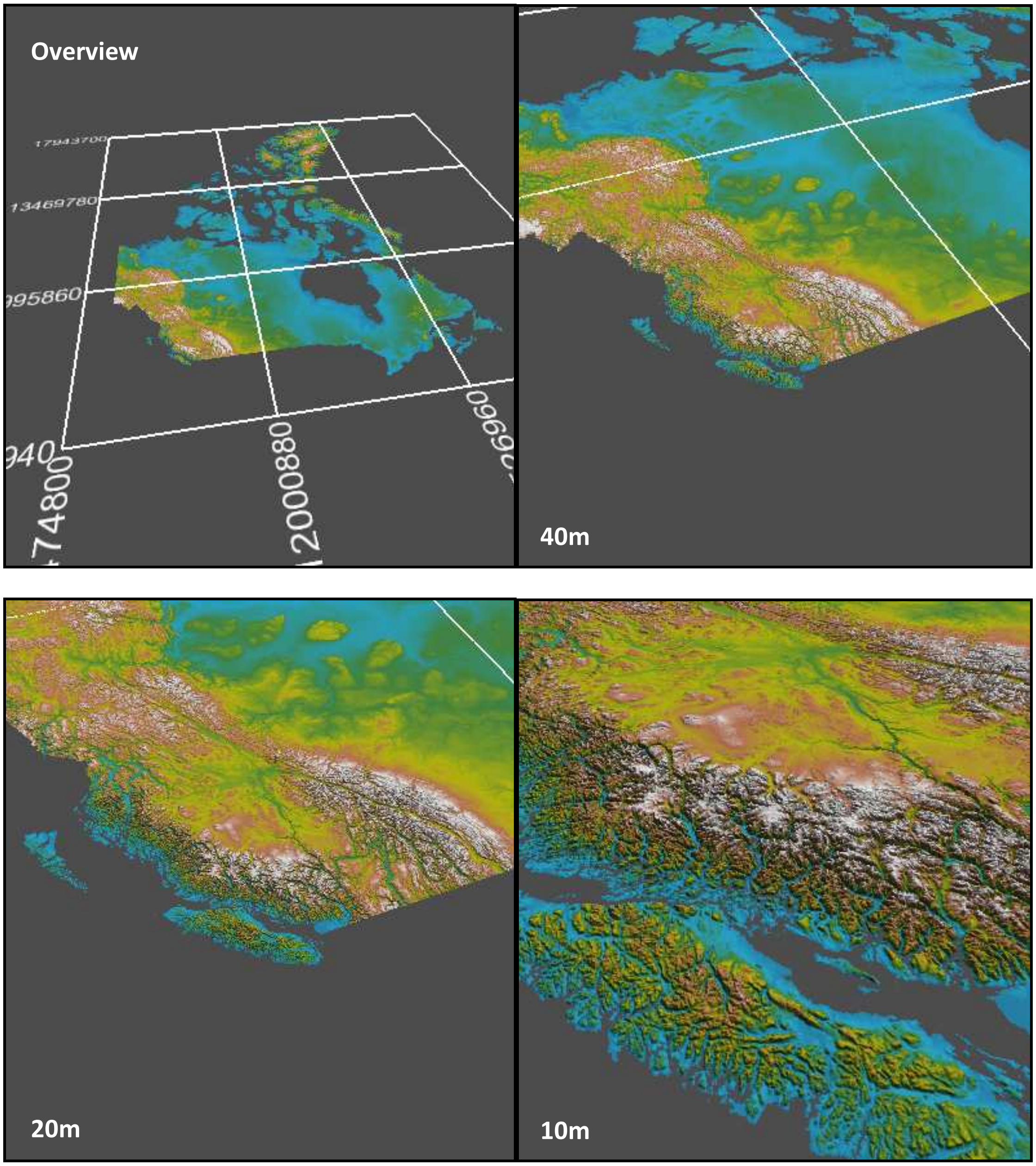

Figure 2: Gridded dataset shown at multiple resolutions.

Gridded Data

Using the CSAR framework, we have implemented a multi-attributed, multi-resolution, geo-referenced grid that can support billions of gridded nodes. The grid at the finest resolution level is partitioned into chunks that store all of the data for a local grid region. Coarser level approximations are generally constructed from the finest resolution data using programmable *updaters*, though each resolution level can be managed independently. Updaters also manage dependencies across attribute bands; therefore, if a node is copied across resolution levels, all of its attributes can follow it. This allows for very flexible update strategies such as *shoal-biased* multi-resolution grids, in which the data at the coarsest level of the pyramid represents the minimum depth in a given area.

Note that, while the independence of the data in each resolution level does allow for very flexible update strategies, it does come at a cost: multi-resolution grids typically require 1.3 to 1.5 times as much storage space as single-resolution grids. The grid structure has been tested on a grid with up to 50 billion nodes.

Point Data

We have also implemented a sophisticated multi-resolution, multi-attributed 3D point cloud that can store hundreds of millions of multi-attributed 3D points. The cloud also supports rapid selection of entire ranges of data to facilitate editing millions of points simultaneously.

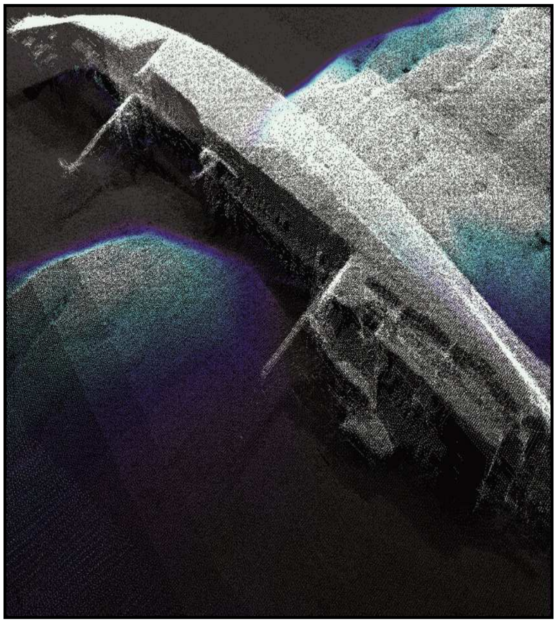

Figure 3: CSAR Point Cloud (image courtesy of UKHO; data collected with a Kongsberg EM710 by Fugro OSAE GmbH on behalf of the UK Maritime and Coastguard Agency)

As with the grid, data is moved across resolution levels using programmable updaters, which also manage the dependencies between the attributes attached to each 3D point. The point cloud has been tested on point sets of 300 million points, though preliminary results suggest that much higher volumes can be stored.

Storage

The CSAR framework is a portable file format for the storage of terabytes of data, along with associated metadata, thereby supporting the concept of a “Field Database.” It was designed to impose minimal constraints on the storage devices it works with. The first file-based implementation—termed the CARIS Spatial Archive—of a storage backend for the framework uses an open-source, embedded database that was extended to handle the large volumes of data stored in the framework's data chunks. This implementation has successfully stored data sets as large as 1 TB, and can store multiple grids and point clouds, as well as any other data structures that may be developed in the future. This implementation also supports pre-fetching of data: a data chunk can be requested without blocking existing processing.

Though the first implementation of the CSAR framework is file-based, the framework can interact with any commercial RDBMS system, other file types, and prototypes using commercial RDBMS systems have already been developed. Translation between the data structures used by those systems and the CSAR frameworks' internal format can be done at the level of the storage layer. This will allow the CSAR framework to support interaction with open storage formats.

Many clients require the scalability, management tools, and robustness afforded by a commercial RDBMS. Oracle Spatial has recently introduced its GeoRaster and Point Cloud structures, two new storage formats for Gridded and Point Cloud data, respectively. These new data formats are similar in overall structure to their CSAR equivalents. By writing new CSAR storage layers, data stored in Oracle GeoRasters or Point Clouds will be translated to and from the equivalent CARIS formats on the fly. This will allow applications to take advantage of the query and I/O mechanisms built into Oracle Spatial to read and write chunks of data, without having to modify interfaces at the application level. Furthermore, because the framework has built-in caching to improve processing speed and is linked to the CARIS visualization and processing technology, the data stored in Oracle Spatial will immediately be viewable in 2D and 3D and can be processed using the tools already built into the Bathy DataBASE.

Data Processing and Management

The CSAR framework has been implemented in HIPS and SIPS and Bathy DataBASE, which are part of the CARIS Ping to Chart workflow. The Ping to Chart workflow has been designed to reduce bathymetry and image data processing times relative to acquisition. The new CSAR framework will support the workflow by further increasing processing efficiency.

HIPS and SIPS

Hydrographic data processing starts with the analysis of pings of raw bathymetry soundings and/or imagery data that are collected during survey acquisition. As part of the CARIS Ping to Chart workflow, gridded products can be created in HIPS and SIPS in order to simplify the very dense data collected during acquisition. The high density raw sounding data, for example, may include millions or even billions of points in a single survey area. The final gridded products created in HIPS and SIPS (or Bathy DataBASE) are called BASE (Bathymetry and Associated Statistical Error) Surfaces.

The most significant advancement for the latest version of HIPS is the incorporation of the CSAR storage framework. The BASE Surfaces created in HIPS are stored in the CSAR framework. There are three main benefits to gridding a hydrographic dataset in the CSAR framework in HIPS.

First, CSAR supports not only bathymetry surfaces but in a future release of HIPS will support imagery data (e.g. side scan or backscatter mosaics), leading to a more integrated workflow overall. For example a user will be able to compare multibeam bathymetry and side scan sonar datasets collected concurrently in order to identify objects (dangers to navigation), and further correlate these objects between the datasets. The faster visualization and processing of data in CSAR will enable more efficient analysis of all hydrographic data collected during a single survey operation. In addition, storing both bathymetry and imagery as a CSAR surface will allow more efficient storage to one database, a “Field Database” used for all hydrographic data collected for multiple surveys.

Second, the gridding process is multi-threaded, allowing full utilization of machine hardware and dramatically increasing processing speed / gridding time.

And third, a grid size “limitation” has not been reached with the new CSAR framework. As mentioned previously, a CSAR grid can contain billions of nodes, saved in a single dataset tested to a size of 1 TB. Because of the CSAR framework design, mainly its ‘multi-resolution’ visualization aspect, it is quick and easy to load, view, and manipulate a very large gridded dataset. This eliminates the need to break up a survey and surfaces into smaller pieces (e.g., by creating small fieldsheets). The only size limitation found with initial testing has been hard drive space.

Bathy DataBASE

The CARIS Bathy DataBASE (BDB) software suite, consisting of a desktop application and a server module, is the next step in the CARIS Ping to Chart workflow. The suite includes tools for the management, compilation, validation, and storage of hydrographic data. Gridded bathymetry surfaces in the CSAR format that were created in HIPS can be easily compiled in BDB, or new BASE Surfaces can be created from existing processed sounding datasets. These data could include numerous large-volume datasets that can all be stored on the BDB server.

Like HIPS and SIPS, BDB incorporates the same functionality with the CSAR framework implemented, including storage of bathymetry and imagery data in CSAR format, multi-threaded processing, and no imposed software limitation reached on dataset size. Furthermore, the storage of multi-resolution 3D point cloud data is supported in addition to gridded data. The point cloud enables the storage of hundreds of millions of multi-attributed 3D points. This provides a more integrated workflow because the same framework is used for storing point *and* gridded data. For example, a user can create new designated soundings on gridded data or create new features directly from nodes *or* points, capturing attributes about them.

The adoption of the CSAR framework into the Bathy DataBASE product allows the product to scale up at the individual dataset level. Previous versions of the BDB software were scalable in terms of number of datasets, but the current release can handle individual datasets of enormous volume. Now it is possible to work with much larger areas as well as denser datasets more efficiently. BDB used to store its gridded and point bathymetry in a proprietary format (HNS and BPS, respectively) that, while very well suited to moderately-sized data sets, did not scale efficiently to very large data sets of the kind that are now being generated using modern sensor systems. The format was also not designed to facilitate network access or remote visualization of large datasets. Now, users can quickly load and view very large datasets over a network connection and/or from a database location on a server.

Below are the results of some initial comparisons of datasets stored in the current HNS format and the same datasets converted into the equivalent CSAR grids.

		HNS Format			**CSAR Format**		
Surface	Size	Open	Overview	Refresh	Open	Overview	Refresh
Surface A - Local	1.6 Gb	135	55	2	3	1	1
Surface A - Network	1.6 Gb	365	55	6	3	1	1
Surface B - Local	630 Mb	25	1	2	3	1	1
Surface B - Network	630 Mb	72	1	2	3	1	1

Table 1: HNS file versus CSAR file drawing times (seconds) on Local machine and across a network connection

Table 1 contains the results of opening and manipulating (overview, refresh) a HNS file as compared to the new CSAR file on a local disk and across a network connection. It is apparent that the CSAR format behaves much more efficiently, especially when it comes to grids of larger sizes. For example, it is apparent that a 1.6 GB surface opens and draws in the order of five times faster in the CSAR format as compared to the current HNS format. We see great dividends over the network as compared to local disk for the HNS format but not for the CSAR format. Note that in both cases (network and local files) open, overview, and refresh times stay constant as file size increases. Also, tests with the new data grid storage structure show that load and overview times are independent of the number of nodes in the grid.

Furthermore, clients using BDB (or HIPS) are no longer limited to breaking their datasets down into manageable pieces either for visualization or data management. Surfaces can be combined over large areas yielding high-volume point cloud or gridded datasets that are seamlessly loaded and drawn on the fly. This allows for the creation of bathymetric products from these large-volume datasets using new high-precision geometries on data in the CSAR format. Products might include bathymetric Electronic Navigational Charts (bENCs), which would complement existing ENCs by allowing the mariner to view high-density bathymetry.

Finally, the CSAR framework in BDB will allow non-bathymetric data formats like NetCDF to be supported. The CSAR point cloud itself can support multiple data values (z) at a single location (x, y). So, the “z” value does not have to be a depth. Therefore, other non-bathymetric data like oceanographic, geological, or meteorological will be supported in the future depending on client needs.

3D Visualization

To take advantage of the multi-resolution structure of our new grid and point storage format, and handle high data volumes, we have reimplemented our 3D visualization system. This new functionality is available in both HIPS and SIPS and Bathy DataBASE. The new system is multi-threaded, and loads data across resolution levels on background threads while the scene is drawing. The viewer always sees some representation of the data on screen, even when the background data is being loaded across a low-bandwidth network connection. This allows the data to be viewed efficiently when accessed from the Bathy DataBASE Server by the BASE Manager client when those systems are connected over a network.

The system also supports dynamic lighting and texturing and other environmental settings that allow the user to easily interact with the data in 3D. As well, vector or raster data can simply be draped across the scene to further aid in feature investigation and identification.

Conclusions

The increasing demand for bathymetric products has also increased demand for simpler, more open, and more efficient mechanisms for warehousing and distributing hydrographic information. To help meet this demand CARIS has overhauled its data storage technology and visualization engines to handle very high volume datasets. This new technology has been directly realized in HIPS and SIPS and the Bathy DataBASE software suite as part of the CARIS Ping to Chart workflow.

The CSAR framework was designed to efficiently read, write, and process large collections of pieces of data known as chunks. The framework is modular and layered, and is storage-system-independent. Using this framework, CARIS has implemented high volume storage structures for both gridded and 3D point data types. These structures are designed for fast data access and spatial queries, and are accompanied by a rebuild visualization engine that is capable of loading chunks of data from storage into the 3D view in the background while the scene is being drawn.

Now that the new CSAR technology has been established as a backbone of several of the CARIS Ping to Chart products, we will continue to deliver innovative solutions that increase data processing and management workflow efficiency.