SI2-SSE:Wavelet-enabled progressive data Access and Storage Protocol (WASP)

Lawrence Frank (lfrank@ucsd.edu) and John Clyne (clyne@ucar.edu)

Problem

Advances in modern digital imaging methods are revolutionizing a wide range of scientific disciplines by facilitating the acquisition of huge amounts of data that allow the visualization, measurement, reconstruction, and archiving of complex, multi-dimensional images. Concurrently, modern computing technologies enable numerical modeling of a broad gamut of scientific phenomena, resulting in the production of vast quantities of numerical data, which are just the starting point for the scientific exploration that modern computational and visualization methods enable. This is particularly true in the biological and geosciences, two seemingly very different disciplines. But these advanced capabilities come at a cost shared by both bio- and geo-sciences: Increasing data size and complexity require more sophisticated methods for data management, analysis and visualization. This proposal, titled **SI2-SSE: Wavelet-enabled progressive data Access and Storage Protocol (WASP)**, provides a common software framework for supporting *a multi-scale progressive data refinement method based upon the representation of the data as a wavelet expansion*[1], and enabling interactive exploration of large data sets for the bio- and geoscience communities.

Progressive Data Access

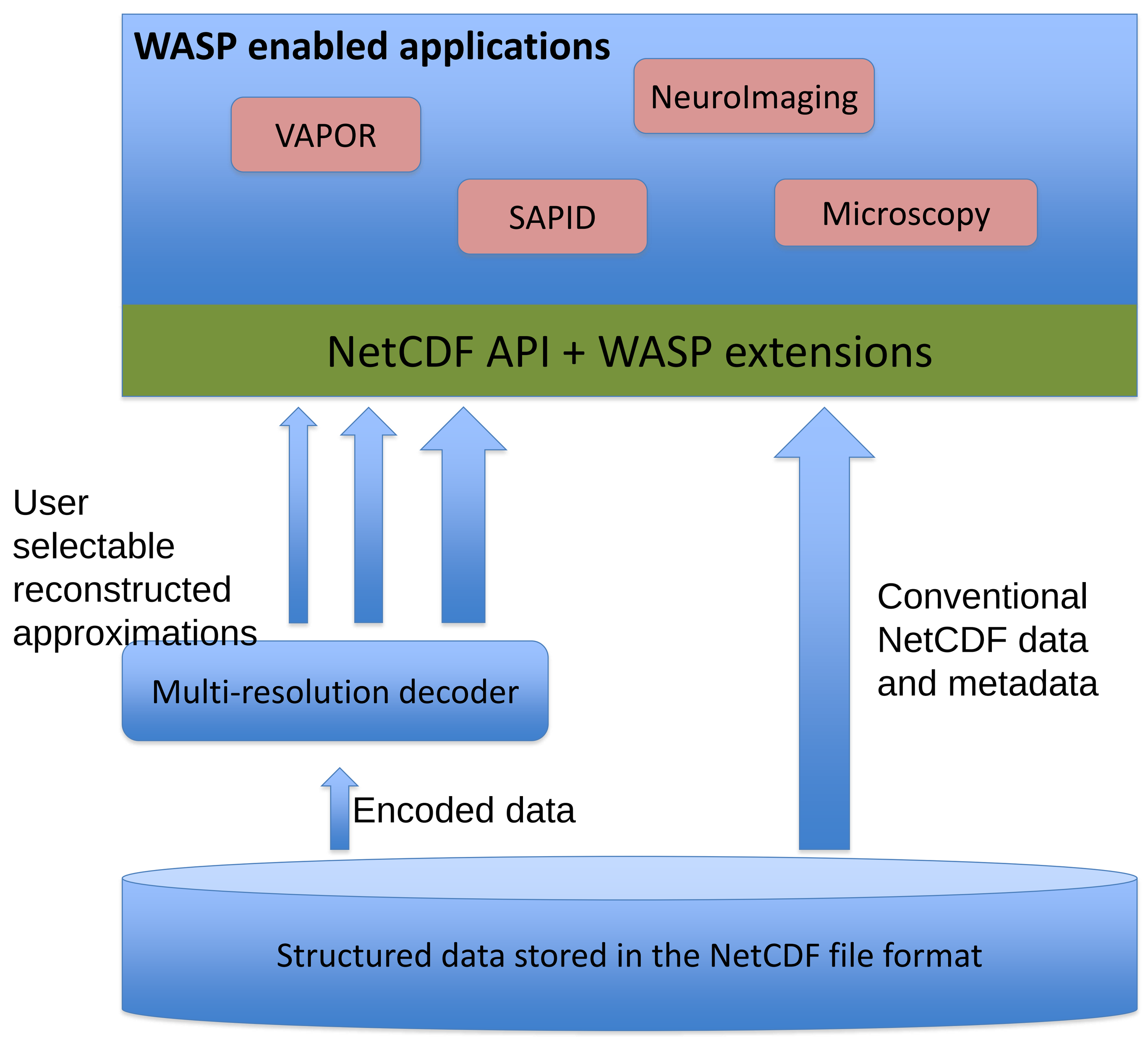
Biological imaging datasets have traditionally been smaller than those in other fields such as meteorology, but advances in digital imaging technologies are closing this gap. Efficient management of large data for visualization, and storage supporting scientific discovery is an increasingly non-trivial problem as the size and complexity of imaging and simulation data increases. This project addresses a critical technical challenge that limits the explorative capabilities and the utility of large data: the implementation of a general progressive access, multi-scale data representation for the efficient handling of large, complex, multi-variate data sets. Progressive data access is the enabling technology behind ubiquitous digital mapping technologies such as Google Maps™; coarsened approximations of high resolution maps are presented when the viewpoint is far away. As the viewer moves closer to an area of interest more and more detail is presented. This project will develop and release an open source C/C++ WASP software toolkit that affords structured scientific data sets, and corresponding applications, similar, yet more sophisticated and powerful, data access capabilities as those employed by consumer mapping display technologies.

WASP project goals

VAPOR (www.vapor.ucar.edu) is an NSF funded geosciences data visualization and analysis package with thousands of users worldwide. VAPOR enables the interactive exploration of the highest resolution numerical simulation outputs arising from climate, weather, and other disciplines. The cornerstone of this ability is an intelligent wavelet based storage format that allows the researcher to progressively coarsen or refine data, as needed, thus trading off speed (interactivity) for fidelity. The goals of the WASP project are to enhance and reuse the first generation VAPOR storage model, both improving its utility for the existing VAPOR user community, and generalizing the software to make it suitable for other science domains, in particular the bioimaging community. Specific objectives are to:

1. **Develop open NetCDF file format conventions for progressive data access using wavelets**
2. **Develop an open source toolkit (API) in support of these storage conventions**
3. **Integrate the new toolkit into the extant VAPOR geoscience visualization tools**
4. **Integrate new the toolkit into the STK biological imaging visualization and analysis tools**

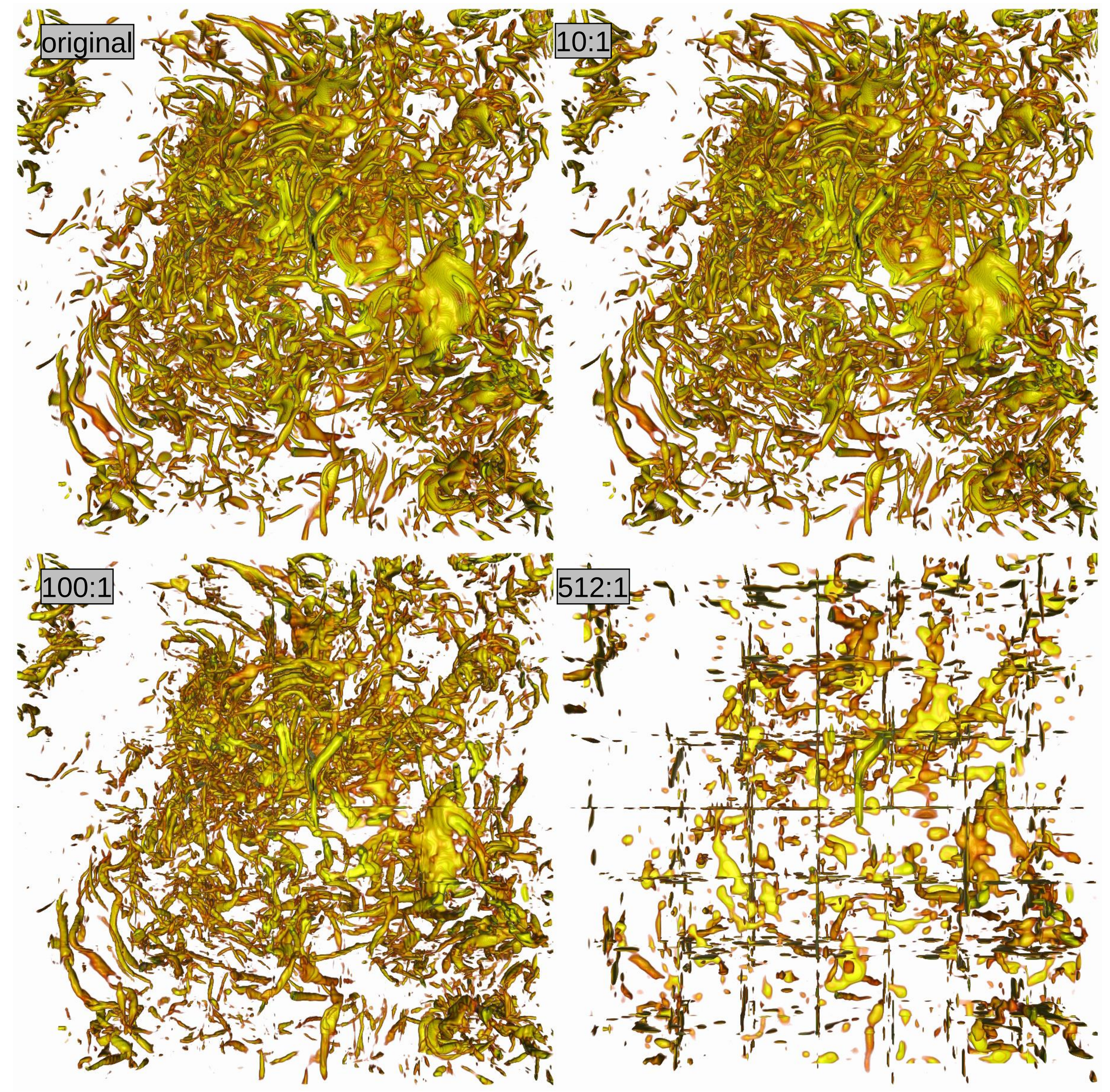
WASP functional diagram



This work is supported by NCAR's Computational & Information Systems Lab. (CISL), UCSD Center for Scientific Computation in Imaging (CSCI) and NSF Grants ABI 1147260, and ACI-1440412, and the Korean Institute for Science and Information Technology (KISTI).

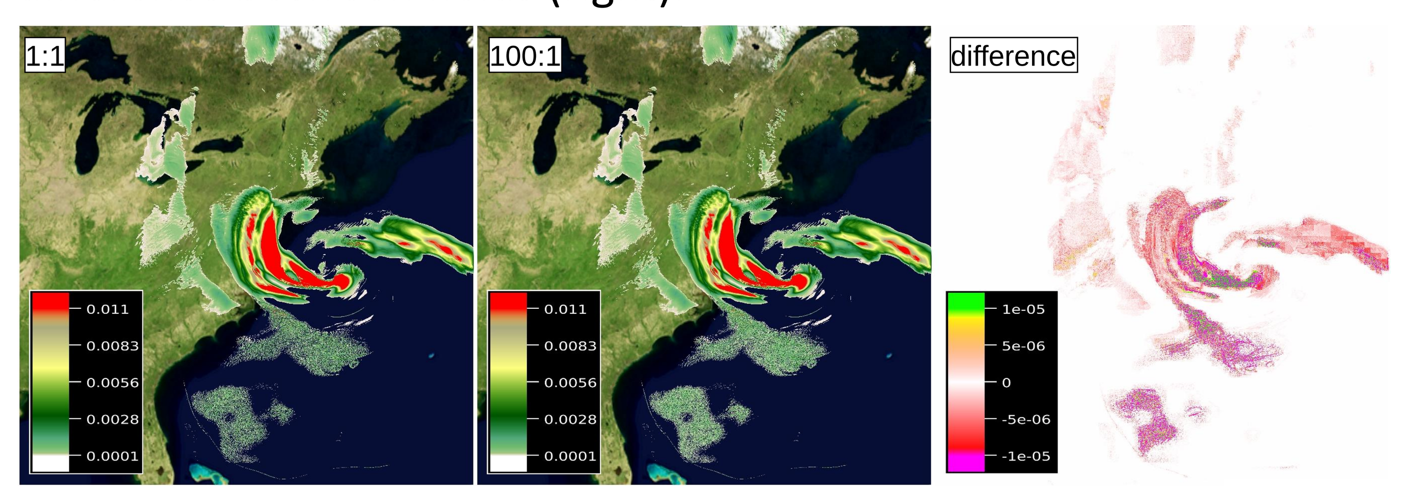
Numerical simulation: particle-turbulence interactions

Here we show qualitative results volume rendering data arising from high resolution numerical simulations that have been compressed by discarding the smallest wavelet coefficients. The 2048^3 simulation explores particle-turbulence interactions in conditions which mimic cumulus cloud cores[3]. Shown is the enstrophy field derived by taking the curl of compressed velocity field components using 6th-order finite differences. Compression rates are indicated in the top-left corner of each sub-figure. The original uncompressed variables occupy 32GBs per time step.



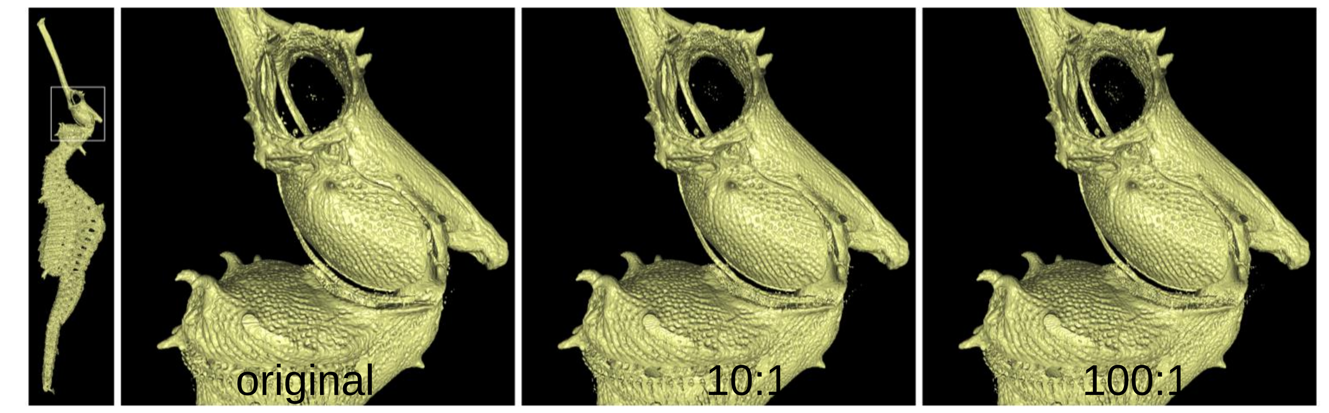
Numerical simulation: Hurricane Sandy

Here we show results for a very high resolution simulation of Hurricane Sandy, computed on a 0.5km grid (~5000x5000x150 grid points)[4]. Rain water mixing ratio summed over a vertical column. Original (left), compressed 100:1 (middle), and difference between fields (right).



Imaging: microCT scan

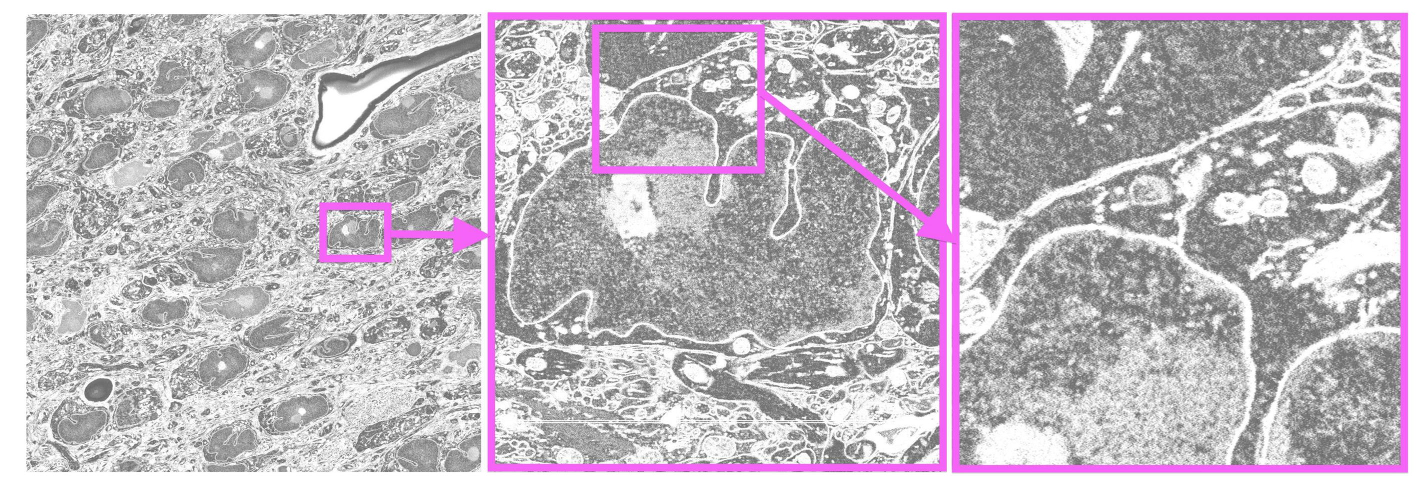
A CT scan at 35mm resolution of a new species of seadragon visualized with progressive data access at full resolution (left) and 1/10th resolution with 1/100th the wavelet coefficients (right). (data courtesy R. Sah, UCSD, specimen courtesy G. Rouse, SIO)



*For the purposes of morphology, those updated images are **indistinguishable** from each other. Every feature of interest remains clearly visible. Even when zooming way in, they are still indistinguishable from each other - the slight differences in pixelation do not impact the visibility of the anatomical features nor their shape.*

Rachel Berquist, zoologist, UCSD

Imaging: microscopy



Zeiss Merlin/Gatan 3View High-Performance SBEM of mouse cerebellum. This figure illustrates the power of progressive data access applied to a massive microscopy data set acquired at NCMIR. The raw data consist of 32,000 by 24,000 by 703 grid points of 16bit quantities, occupying over 1 Terabyte of storage and thus virtually impossible to interactively explore in their raw form. However, using prototype progressive data access techniques an isolated region-of-interest can be rapidly identified and isolated using a coarsened representation, and then examined in detail using the highest fidelity data available, all at highly interactive speeds. (data courtesy M. Ellisman, UCSD)

References

1. S. Mallat. A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way, 2008.
2. J. Clyne and A. Norton. Progressive Data Access for Regular Grids, in High Performance Visualization, 2012
3. Data source: Peter J. Ireland and Lance R. Collins (Cornell University)
4. Data source: Peter Johnsen (Cray), Mel Shapiro (NCAR), and Tom Galerno (NCAR)