Scientific Data Compression

Lawrence Ibarria†, Peter Lindstrom‡, Jarek Rossignac†

†Georgia Institute of Technology ‡Lawrence Livermore National Laboratory,

Scientific computing generates huge loads of data from complex simulations. We present an ongoing study about storing structured floating point data compactly and efficiently without any loss. Our method compresses floating point-based structured grids using a 3D progressive prediction, and a specialized floating point coder.

Problem:

- **Scientific simulations create 3D and 4D grids that take several TB.**
- **Requirements: fast view, reduced storage, fast processing.**
- **Zip compression does not transform the data and has suboptimal results.**
- **State of the Art: Lossless floating point compression using Lorenzo Prediction gives good compression results but is not progressive.**

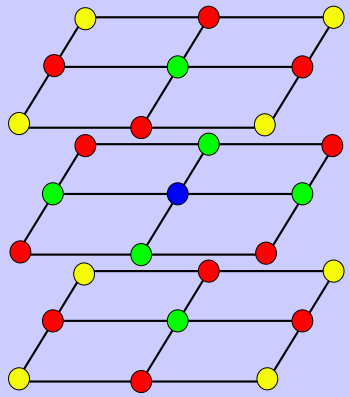
Method:

Prediction is applied progressively onto the data. Data points are classified into volume, face and edge points.

- **Edge points are predicted using a cubic polynomial fitting.**
- **Face points are predicted using the Radial 2D predictor.**
- **Volume points are predicted using the Radial 3D predictor.**

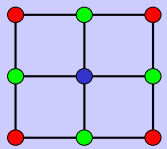
Different probability models are used per predictor and level.

Radial 3D



● = 3*(● - ●) +●

Radial 2D

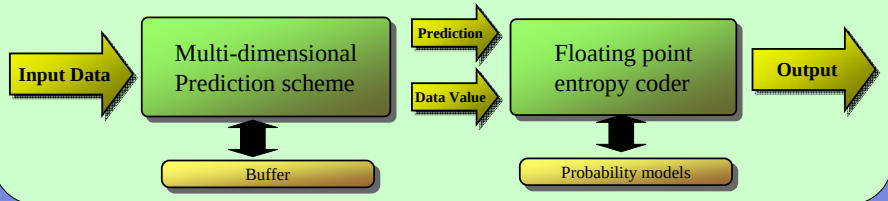


● = 2*● -●

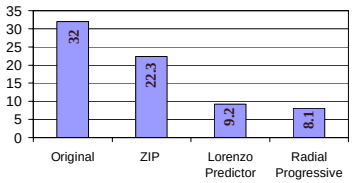
Entropy coders

Entropy coders are probability based. They achieve the optimal compression when the probability distribution is biased, such when a few symbols occur much more frequently than all other symbols.

Overview



Results:



Radial is progressive and improves the best result in a 13% average (up to a 30%). We achieve a 4:1 lossless compression.

Discussion:

The mask of the predictor determines the assumed smoothness of the data. We use a conservative small set of masks. There is room for improvement using variable sized masks, as well as predictors of variable degree. Predictors that have a 3D mask for all non-border points have not yet provided comparable compression ratios.

Predictors:

Predictors can be classified as interpolating and extrapolating. The Lorenzo predictor is extrapolating and uses a single resolution scheme. Radial is an interpolating predictor that uses a progressive approach. The main differences are:

-Radial has better accuracy
-Radial does not use the same predictor for all non-border points (lower dimensional predictors for edge or face points).
-The progressive approach makes larger errors in higher levels.

This work was performed under the auspices of the U.S. Department of Energy by University of California Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48.

UCRL