SALSA PROJECTS

http://salsahpc.indiana.edu/

# Scalable Dimension Reduction for Large Abstract Data Visualization

**Seung-Hee Bae, Judy Qiu, and Geoffrey Fox**

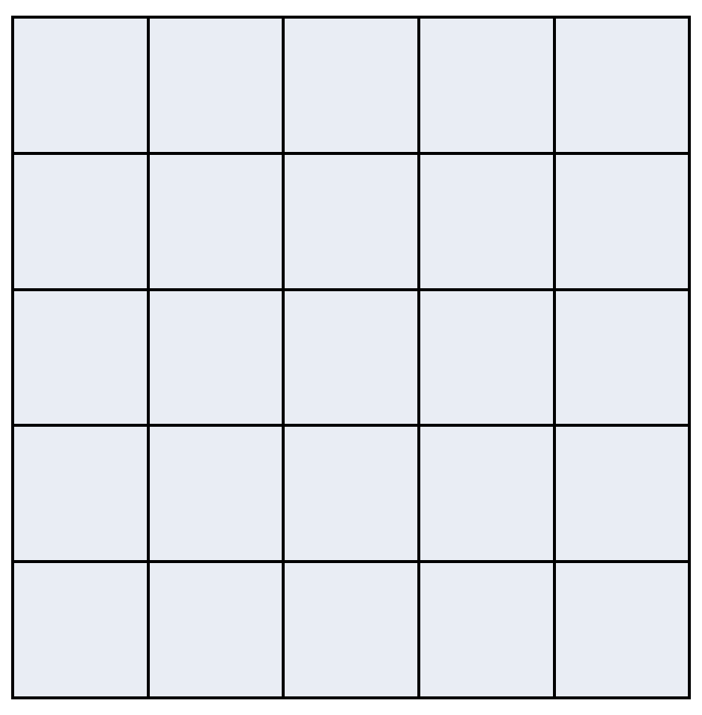
{sebae, xqiu, gcf}@indiana.edu

***Abstract -*** The ability to browse vast amounts of scientific data is critical to facilitate science discovery. High performance Multidimensional Scaling (MDS) algorithm makes it a reality by reducing dimensions so that scientists can gain insight into data set from a 3D visualization space. As multidimensional scaling requires quadratics order of physical memory and computation, a major challenge is to design and implement parallel MDS algorithms that can run on multicore clusters for millions of data points. Bases on our early work of parallel SMACOF algorithm, the authors have developed an interpolated approach, majorizing interpolation MDS (MIMDS). It utilizes the known mapping from a subset of given in-sample data to effectively reduce computational complexity with minor cost of approximation. MI-MDS makes it possible to process huge data set with modest amounts of computation and memory requirements. Our experimental results show that the quality of interpolated mapping is comparable to that of the original SMACOF in a million chemical compounds data, where we construct a configuration of over two-million out-of-sample data into target dimension space.

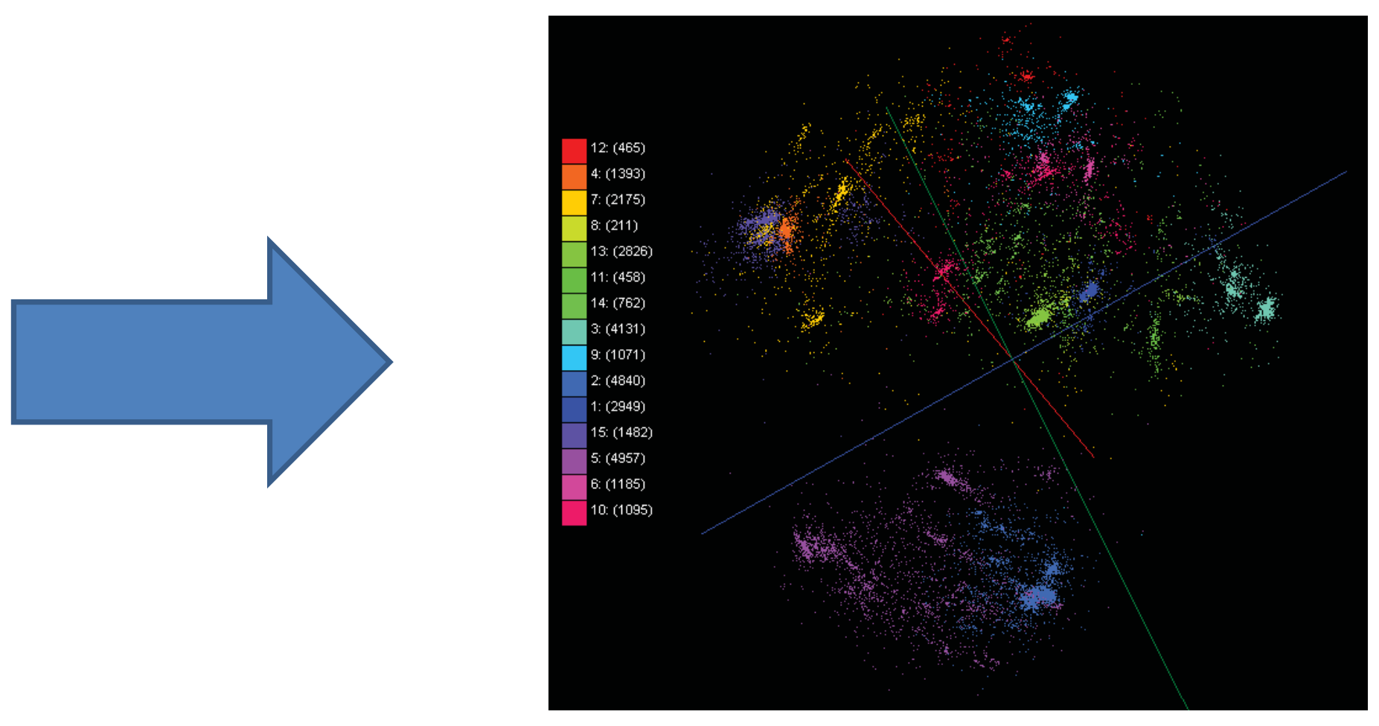
*Keywords*-**Multidimensional Scaling; Parallelism; Interpolation**

## Multidimensional Scaling (MDS)

**Multidimensional scaling (MDS)** is a general term for techniques of constructing a mapping for generally high-dimensional data into a target dimension with respect to the given pairwise proximity information. Mostly, MDS is used for achieving dimension reduction to visualize high-dimensional or abstract data into Euclidean low-dimensional space, i.e. 2D or 3D.



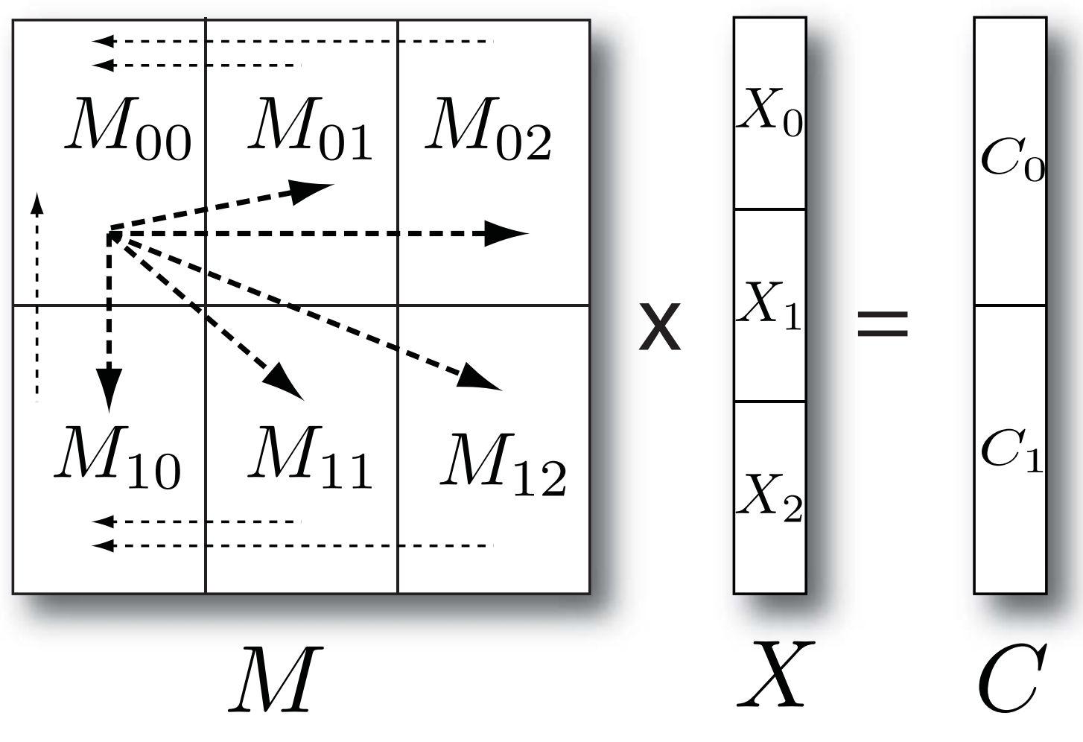
**Pairwise Dissimilarity Data**



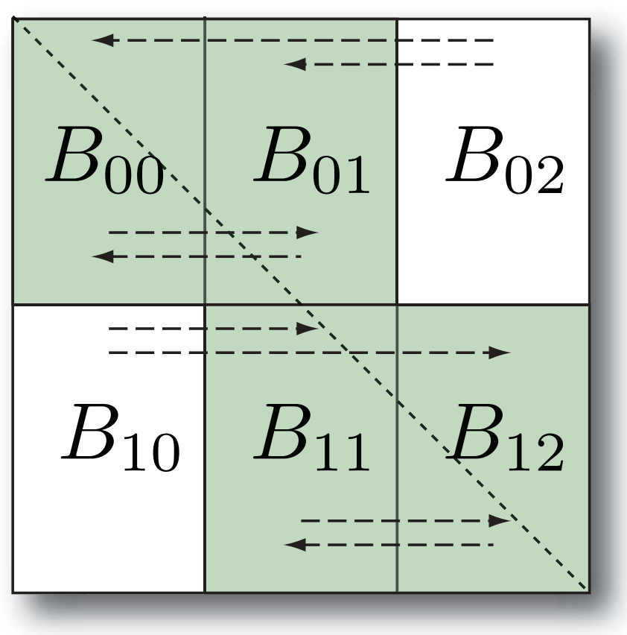
**Mapping in Target Dimension**

## Parallel MDS via MPI

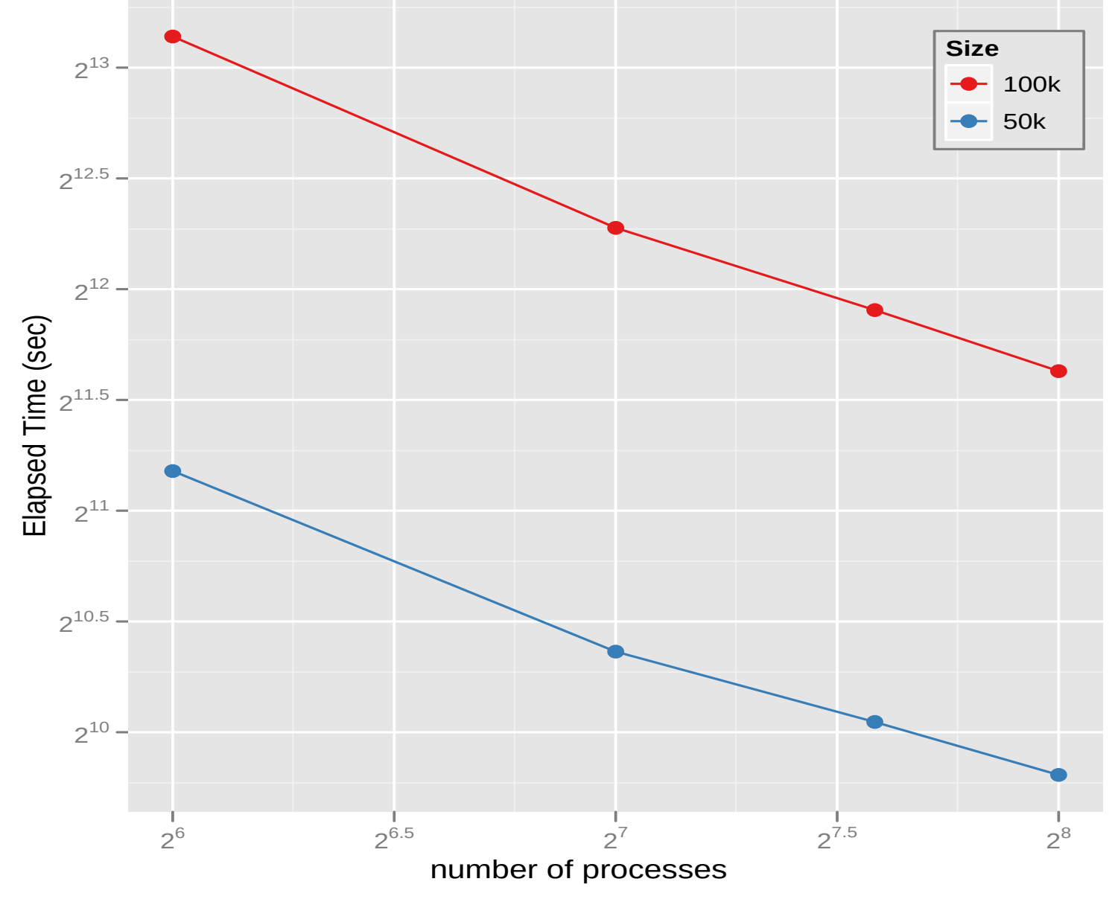
Implement an MDS algorithm called SMACOF, which consists of **matrix multiplications** and **updating status data**, ***in parallel via MPI***.



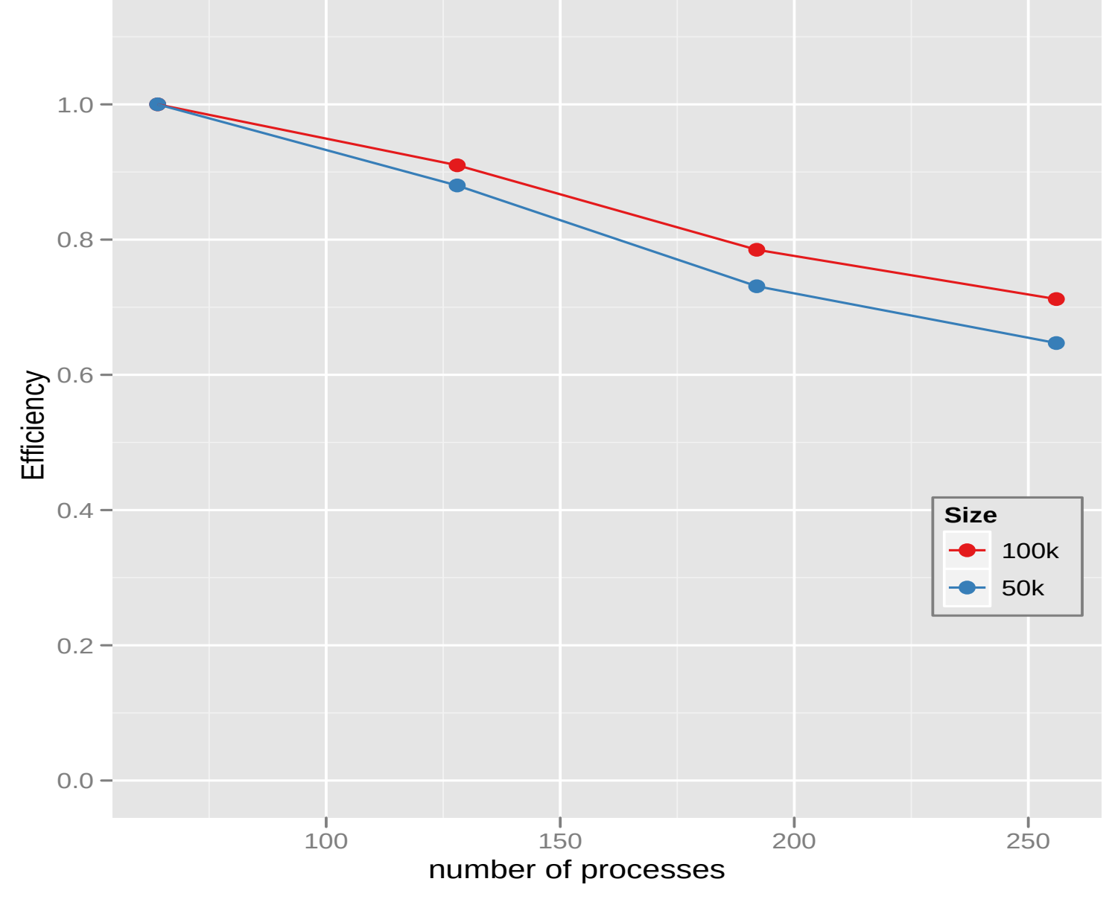
**Parallel Matrix Multiplication**



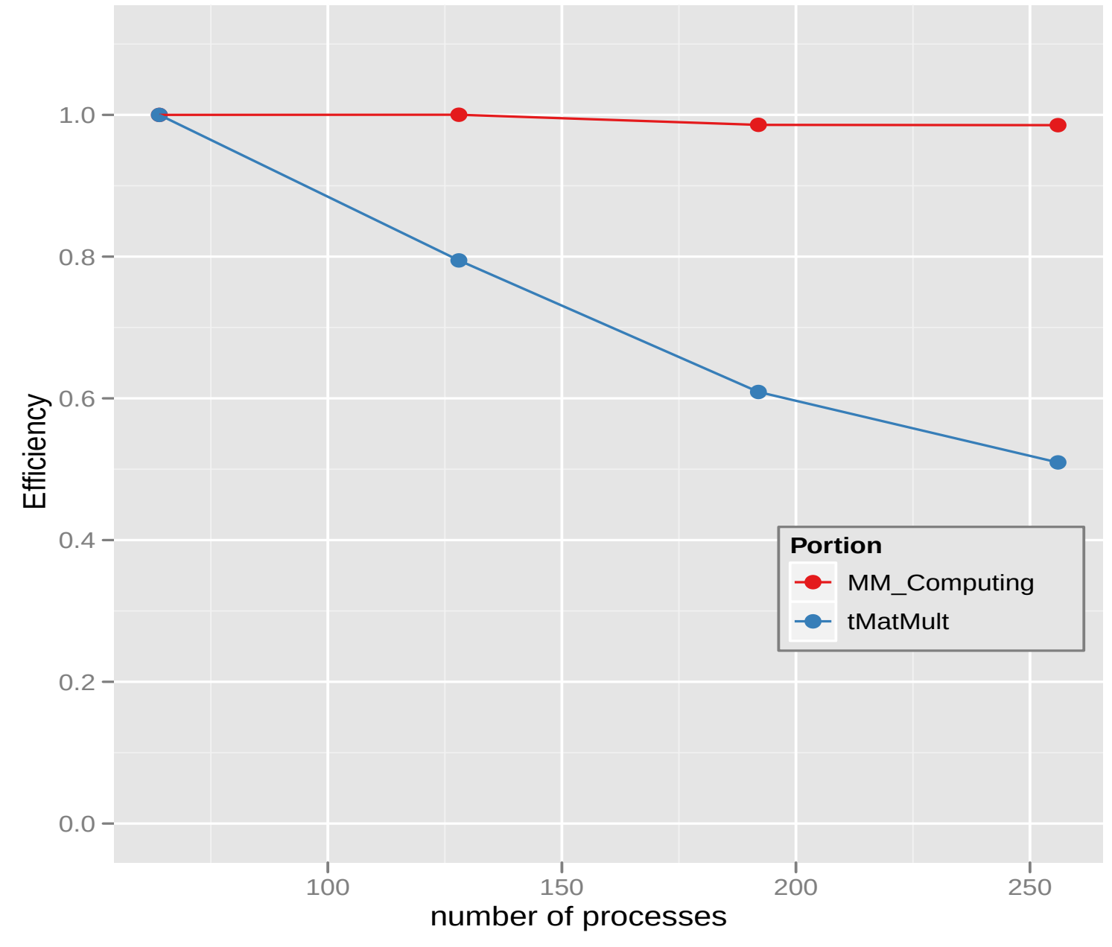
**Parallel data updating**



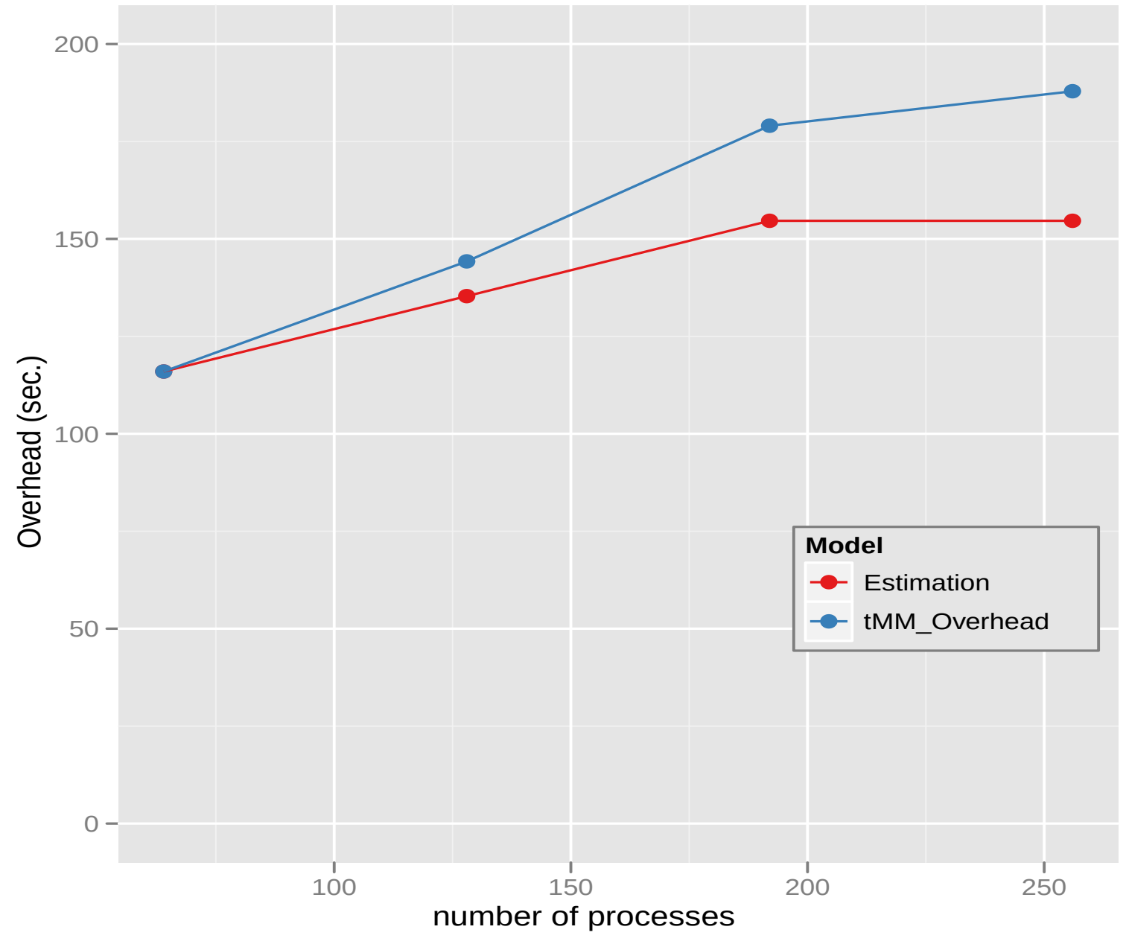
**Parallel Runtime of Large Scale Data**



**Parallel Efficiency of Large Scale Data**

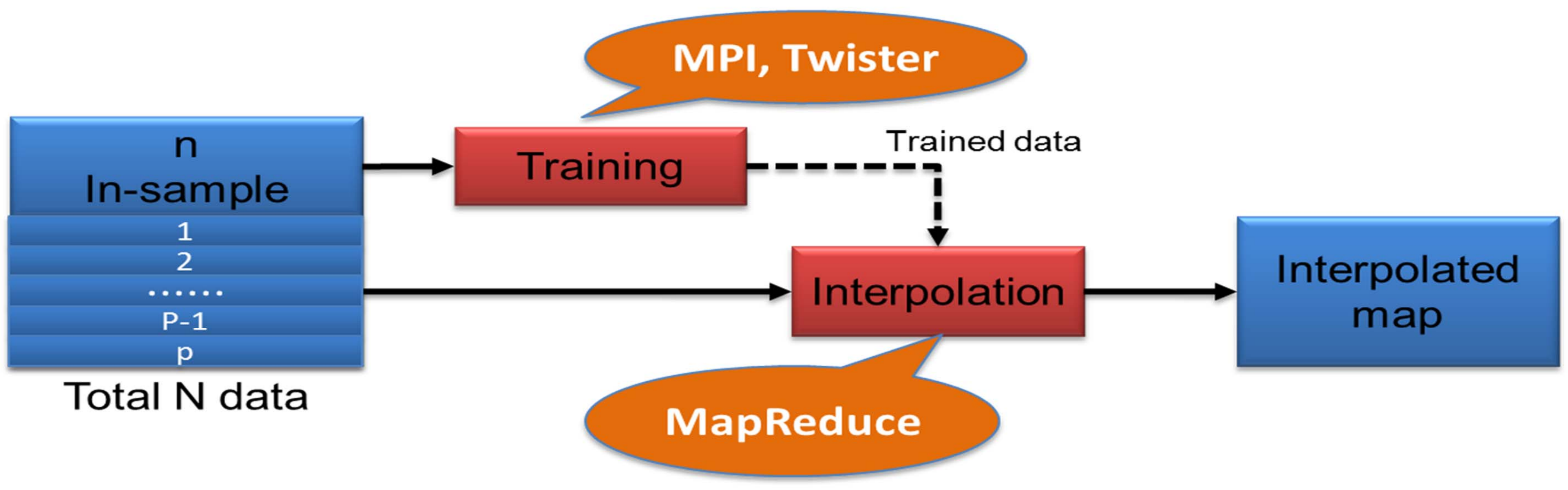


**Efficiency of tMatMult and tMM_Comp**

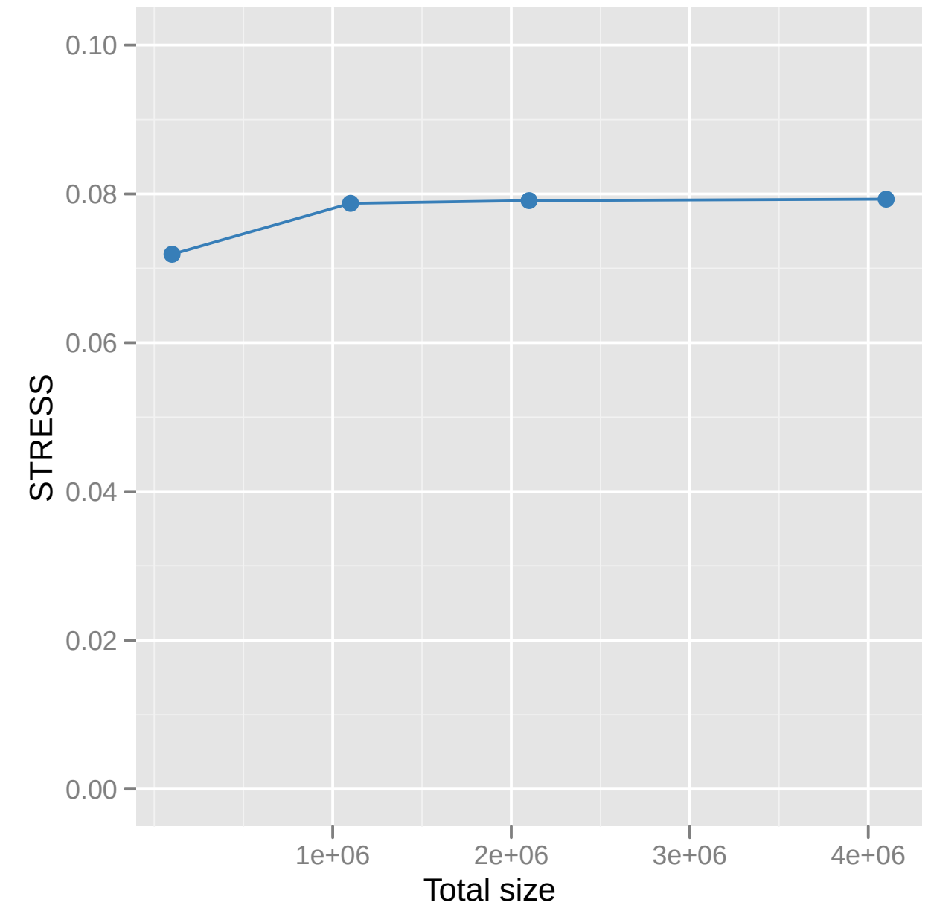


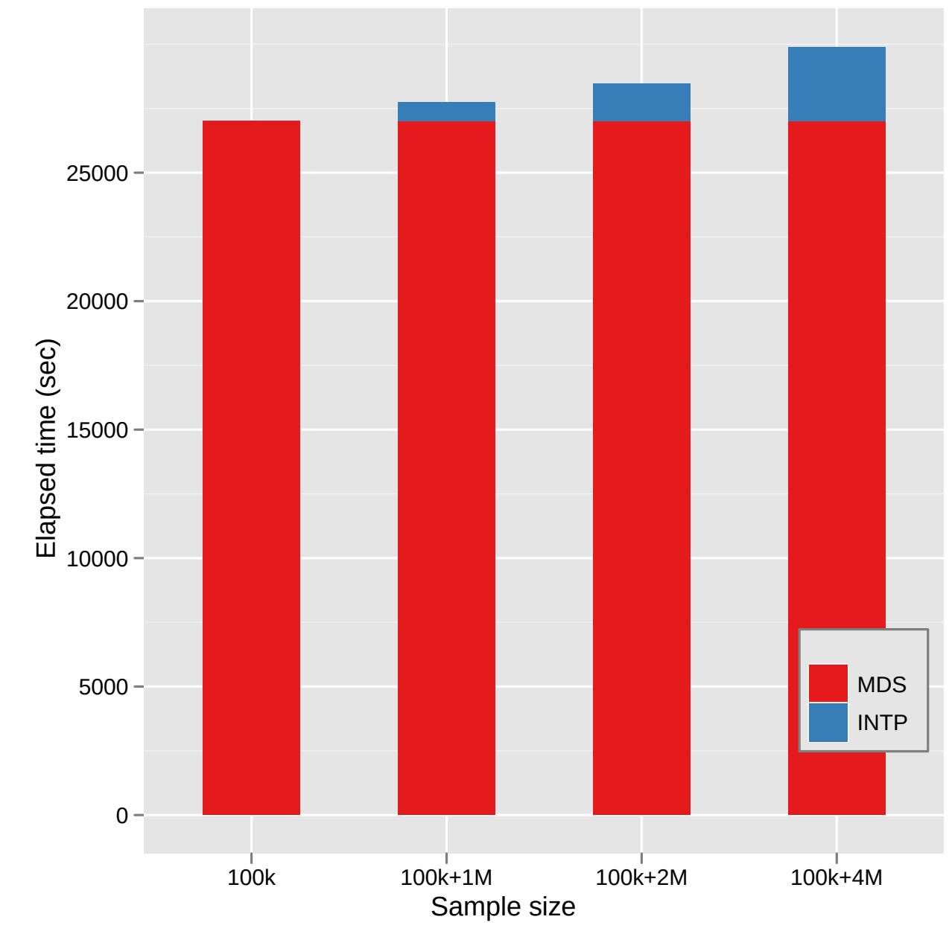
**Overhead of tMatMult and Estimation**

## Parallel Interpolation Approach to MDS (MI-MDS)



**Large scale dimension reduction pipeline**





**Quality Comparison of Large Scale MI-MDS**

**Large Scale MI-MDS Running Time**

| 1M | 2M | 4M |
|---|---|---|
| 731.1567 | 1449.1683 | 2895.3414 |

### Benefits of Interpolation Approach

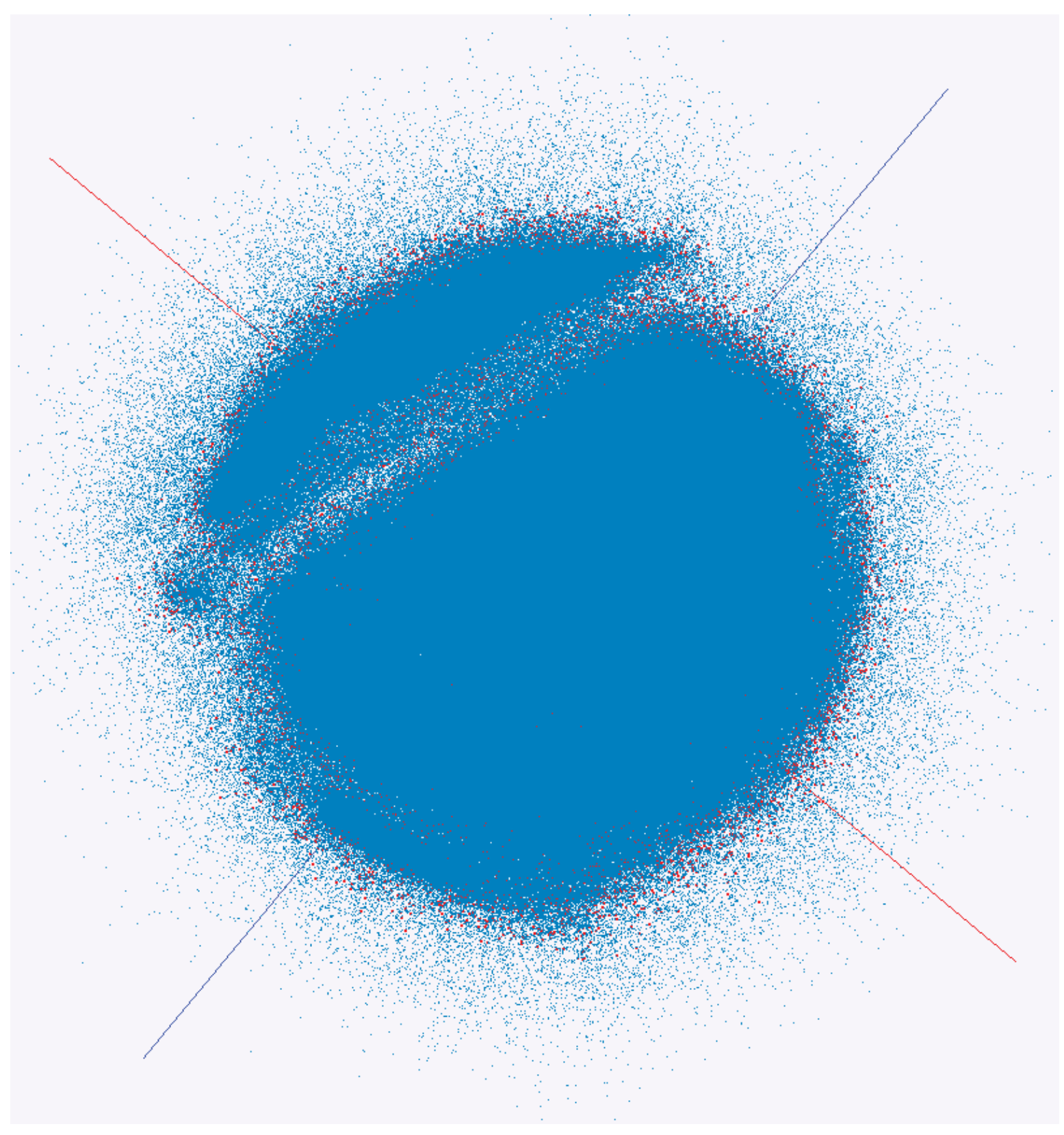
❑ Reduce the resource requirements.

Memory: O($N^2$) → O(n)
Computing: O($N^2$) → O(nM)

❑ Pleasingly Parallel Application.

❑ Mapping quality is good enough.

❑ Extend computational capacity of MDS algorithm into millions of points.



**2M mappings via MI-MDS**

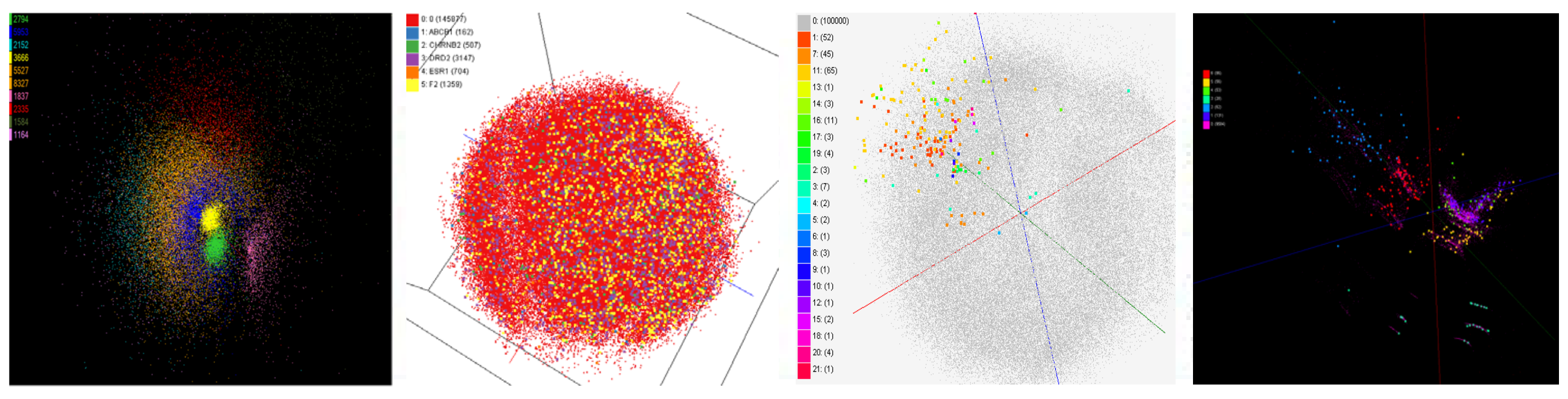
### Related Publications

1. **Seung-Hee Bae**, Judy Qiu, and Geoffrey Fox, "**Multidimensional Scaling by Deterministic Annealing with Iterative Majorizing Algorithm**," in *e-Science 2010.*
2. **Seung-Hee Bae**, Jong Youl Choi, Judy Qiu, and Geoffrey Fox, "**Dimension Reduction and Visualization of Large High-Dimensional Data via Interpolation**," in *HPDC 2010*.
3. Jong Youl Choi, **Seung-Hee Bae**, Judy Qiu, and Geoffrey Fox, "**High Performance Dimension Reduction and Visualization for Large High-Dimensional Data Analysis**," in *CCGrid 2010*.

### Acknowledgement

This work is partially funded by National Institutes of Health grant 1RC2HG005806-01 and Microsoft Research. We would like to thank Prof. Wild and Dr. Zhu for providing chemical compounds data.

PERVASIVE TECHNOLOGY INSTITUTE
INDIANA UNIVERSITY



SALSA HPC