

# Parallel Data Mining from Multicore to Cloudy Grids

Geoffrey FOX<sup>a,b,1</sup>, Seung-Hee BAE<sup>b</sup>, Jaliya EKANAYAKE<sup>b</sup>, Xiaohong QIU<sup>c</sup>, and Huapeng YUAN<sup>b</sup>

<sup>a</sup>*Informatics Department, Indiana University 919 E. 10th Street Bloomington, IN 47408 USA*

<sup>b</sup>*Computer Science Department and Community Grids Laboratory, Indiana University 501 N. Morton St., Suite 224, Bloomington IN 47404 USA*

<sup>c</sup>*UITS Research Technologies, Indiana University, 501 N. Morton St., Suite 211, Bloomington, IN 47404*

**Abstract.** We describe a suite of data mining tools that cover clustering, information retrieval and the mapping of high dimensional data to low dimensions for visualization. Preliminary applications are given to particle physics, bioinformatics and medical informatics. The data vary in dimension from low (2-20), high (thousands) to undefined (sequences with dissimilarities but not vectors defined). We use deterministic annealing to provide more robust algorithms that are relatively insensitive to local minima. We discuss the algorithm structure and their mapping to parallel architectures of different types and look at the performance of the algorithms on three classes of system; multicore, cluster and Grid using a MapReduce style algorithm. Each approach is suitable in different application scenarios. We stress that data analysis/mining of large datasets can be a supercomputer application.

**Keywords.** MPI, MapReduce, CCR, Performance, Clustering, Multidimensional Scaling

## Introduction

Computation and data intensive scientific data analyses are increasingly prevalent. In the near future, data volumes processed by many applications will routinely cross the peta-scale threshold, which would in turn increase the computational requirements. Efficient parallel/concurrent algorithms and implementation techniques are the key to meeting the scalability and performance requirements entailed in such scientific data analyses. Most of these analyses can be thought of as a Single Program Multiple Data (SPMD) [1] algorithms or a collection thereof. These SPMDs can be implemented using different parallelization techniques such as threads, MPI [2], MapReduce [3], and mash-up [4] or workflow technologies [5] yielding different performance and usability characteristics. In some fields like particle physics, parallel data analysis is already commonplace and indeed essential. In others such as biology, data volumes are still such that much of the work can be performed on sequential machines linked together by workflow systems such as Taverna [6]. The parallelism currently exploited is usually the “almost embarrassingly parallel” style illustrated by the independent events in particle physics or the independent documents of information retrieval – these lead

<sup>1</sup> Corresponding Author: Geoffrey Fox; E-mail: gcf@grids.ucs.indiana.edu

to independent “maps” (processing) which are followed by a reduction to give histograms in particle physics or aggregated queries in web searches. MapReduce is a cloud technology that was developed from the data analysis model of the information retrieval field and here we combine this cloud technique with traditional parallel computing ideas. The excellent quality of service (QoS) and ease of programming provided by the MapReduce programming model is attractive for this type of data processing problem. However, the architectural and performance limitations of the current MapReduce architectures make their use questionable for many applications. These include many machine learning algorithms [7, 8] such as those discussed in this paper which need iterative closely coupled computations. Our results find poor results for MapReduce on many traditional parallel applications with an iterative structure in disagreement with earlier papers [7]. In section 2 we compare various versions of this data intensive programming model with other implementations for both closely and loosely coupled problems. However, the more general workflow or dataflow paradigm (which is seen in Dryad [9] and other MapReduce extensions) is always valuable. In sections 3 and 4 we turn to some data mining algorithms that require parallel implementations for large data sets; interesting both sections see algorithms that scale like  $N^2$  ( $N$  is dataset size) and use full matrix operations.

**Table 1.** Hardware and software configurations of the clusters used for testing.

Ref	Cluster Name	# Nodes	CPU	L2 Cache Memory	Operating System
A	Barcelona (4 core Head Node)	1	1 AMD Quad Core Opteron 2356 2.3GHz	2x1MB 8 GB	Windows Server HPC Edition (Service Pack 1)
B	Barcelona (8 core Compute Node)	4	2 AMD Quad Core Opteron 2356 2.3 GHz	4x512K 16GB	Windows Server 2003 Enterprise x64 bit Edition
C	Barcelona (16 core Compute Node)	2	4 AMD Quad Core Opteron 8356 2.3GHz	4x512K 16 GB	Windows Server HPC Edition (Service Pack 1)
D	Barcelona (24 core Compute Node)	1	4 Intel Six Core Xeon E7450 2.4GHz	12 M 48GB	Windows Server HPC Edition (Service Pack 1)
E	Madrid (4 core Head Node)	1	1 AMD Quad Core Opteron 2356 2.3GHz	2x1MB 8 GB	Windows Server HPC Edition (Service Pack 1)
F	Madrid (16 core Compute Node)	8 (128 cores)	4 AMD Quad Core Opteron 8356 2.3GHz	4x512K 16 GB	Windows Server HPC Edition (Service Pack 1)
G	Gridfarm 8 core	8	2 Quad core Intel Xeon E5345 2.3GHz	4x1MB 8GB	Red Hat Enterprise Linux 4
H	IU Quarry 8 core	112	2 Quad-core Intel Xeon 5335 2.00GHz	4x4MB, 8 GB	Red Hat Enterprise Linux 4
I	Tempest (24 core Compute Node) Infiniband	32 (768 cores)	4 Intel Six Core Xeon E7450 2.4GHz	12 M 48 GB	Windows Server HPC Edition (Service Pack 1)

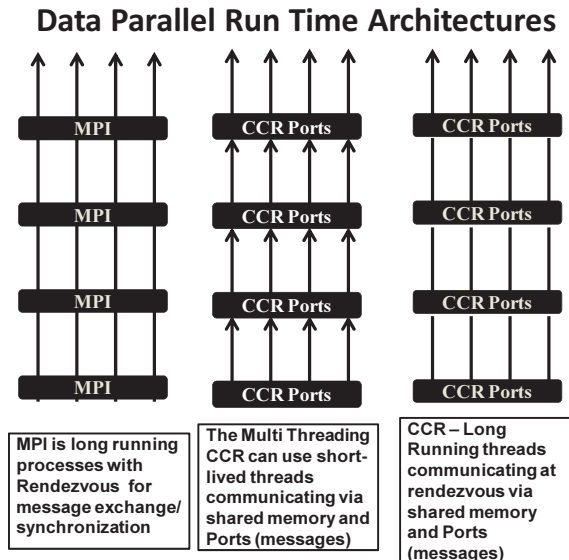
Our algorithms are parallel MDS (Multi dimensional scaling) [10] and clustering. The latter has been discussed earlier by us [11-15] but here we extend our results to larger systems – single workstations with 16 and 24 cores and a 128 core (8 nodes with 16 cores each) cluster described in table 1. Further we study a significantly different clustering approach that only uses pairwise distances (dissimilarities between points) and so can be applied to cases where vectors are not easily available. This is common in biology where sequences can have mutual distances determined by BLAST like algorithms but will often not have a vector representation. Our MDS algorithm also only uses pairwise distances and so it and the new clustering method can be applied broadly. Both our original vector-based (VECDA) and the new pairwise distance (PWDA) clustering algorithms use deterministic annealing to obtain robust results. VECDA was introduced by Rose and Fox almost 20 years ago [16] and has obtained good results [17] and there is no clearly better clustering approach. The pairwise extension PWDA was developed by Hofmann and Buhmann [18] around 10 years ago but does not seem to have been used in spite of its attractive features – robustness and applicability to data without vector representation. We complete the algorithm and present a parallel implementation in this paper.

As seen in table 1, we use both Linux and Windows platforms in our multicore and our work uses a mix of C#, C++ and Java. Our results study three variants of MapReduce, threads and MPI. The algorithms are applied across a mix of paradigms to study the different performance characteristics.

### 1. Choices in Messaging Runtime

The focus of this paper will be comparison of runtime environments for both parallel and distributed systems. There are successful workflow languages which underlies the

approach of the SALSA project [15] which is to use workflow technologies – defined as orchestration languages for distributed computing for the coarse grain functional components of parallel computing with dedicated low level direct parallelism of kernels. At the run time level, there is much similarity between parallel and distributed run times to the extent that both support messaging but with different properties. Some of the choices are shown in figure 1 and differ by both hardware and programming models. The hardware support of parallelism/concurrency varies from shared memory multicore, closely coupled (e.g. Infiniband



**Figure 1(a).** First three of seven different combinations of processes/threads and intercommunication mechanisms discussed in the text

connected) clusters, and the higher latency and possibly lower bandwidth distributed systems. The coordination (communication and synchronization) of the different execution units vary from threads (with shared memory on cores); MPI between cores or nodes of a cluster; workflow or mash-ups linking services together; the new generation of cloud data intensive programming systems typified by Hadoop [19] (implementing MapReduce) and Dryad. These can be considered as the workflow systems of the information retrieval industry but are of general interest as they support parallel analysis of large datasets. As illustrated in the figure the execution units vary from threads to processes and can be short running or long lived.

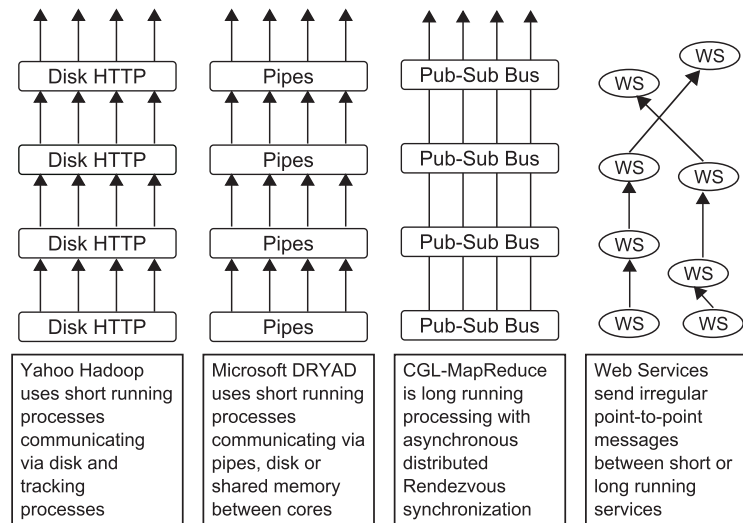


Figure 1(b). Last four of seven different combinations of processes/threads and intercommunication mechanisms discussed in the text

Short running threads can be spawned up in the context of persistent data in memory and so have modest overhead seen in section 4. Short running processes in the spirit of stateless services are seen in Dryad and Hadoop and due to the distributed memory can have substantially higher overhead than long running processes which are coordinated by rendezvous messaging as later do not need to communicate large amounts of data – just the smaller change information needed. The importance of this is emphasized in figure 2 showing data intensive processing passing through multiple “map” (each map is for example a particular data analysis or filtering operation) and “reduce” operations that gather together the results of different map instances corresponding typically to a data parallel break up of an algorithm. The figure notes two important patterns

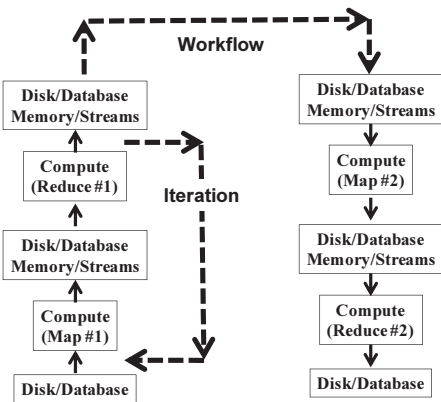


Figure 2: Data Intensive Iteration and Workflow

Workflow

Iteration

a) **Iteration** where results of one stage are iterated many times. This is seen in the “Expectation Maximization” EM steps in the later sections where for clustering and MDS, thousands of iterations are needed. This is typical of most MPI style algorithms.

b) **Pipelining** where results of one stage are forwarded to another; this is functional parallelism typical of workflow applications. In applications of this paper we implement a three stage pipeline:

Data (from disk) → Clustering → Dimension Reduction (MDS) → Visualization

Each of the first two stages is parallel and one can break up the compute and reduce modules of figure 2 into parallel components as shown in figure 3. There is an

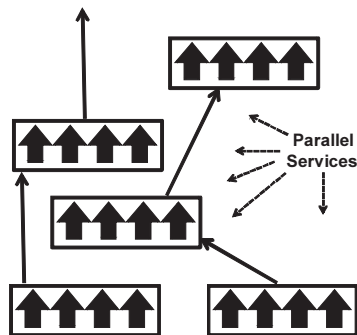


Figure 3: Workflow of Parallel Services

important ambiguity in parallel/distributed programming models/runtimes that both the parallel MPI style parallelism and the distributed Hadoop/Dryad/ Web Service/Workflow models are implemented by messaging. Thus the same software can in fact be used for all the decompositions seen in figures 1-3. Thread coordination can avoid messaging but even here messaging can be attractive as it avoids many of the error scenarios seen in shared memory thread synchronization. The CCR threading [8-11, 20-21] used in this paper is coordinated by reading and

writing messages to ports. As a further example of runtimes crossing different application characteristics, MPI has often been used in Grid (distributed) applications with MPICH-G popular here. Again the paper of Chu [7] noted that the MapReduce approach can be used in many machine learning algorithms and one of our data mining algorithms VECDA only uses map and reduce operations (it does not need send or receive MPI operations). We will show in this paper that MPI gives excellent performance and ease of programming for MapReduce as it has elegant support for general reductions although it does not have the fault tolerance and flexibility of Hadoop or Dryad. Further MPI is designed for the “owner-computes” rule of SPMD – if a given datum is stored in a compute node’s memory, that node’s CPU computes (evolves or analyzes) it. Hadoop and Dryad combine this idea with the notion of “taking the computing to the data”. This leads to the generalized “owner stores and computes” rule or crudely that a file (disk or database) is assigned a compute node that analyzes (in parallel with nodes assigned different files) the data on its file. Future scientific programming models must clearly capture this concept.

## 2. Data Intensive Workflow Paradigms

In this section, we will present an architecture and a prototype implementation of a new programming model that can be applied to most composable class of applications with various program/data flow models, by combining the MapReduce and data streaming techniques and compare its performance with other parallel programming runtimes such as MPI, and the cloud technologies Hadoop and Dryad.

MapReduce is a parallel programming technique derived from the functional programming concepts and proposed by Google for large-scale data processing in a

distributed computing environment. The *map* and *reduce* programming constructs offered by MapReduce model is a limited subset of programming constructs provided by the classical distributed parallel programming models such as MPI. However, our current experimental results highlight that many problems can be implemented using MapReduce style by adopting slightly different parallel algorithms compared to the algorithms used in MPI, yet achieve similar performance to MPI for appropriately large problems. A major advantage of the MapReduce programming model is that the easiness in providing various quality of services. Google and Hadoop both provide MapReduce runtimes with fault tolerance and dynamic flexibility support.

Dryad is a distributed execution engine for coarse grain data parallel applications. It combines the MapReduce programming style with dataflow graphs to solve the computation tasks. Dryad considers computation tasks as directed acyclic graph (DAG)s where the vertices represent computation tasks –typically, sequential programs with no thread creation or locking, and the edges as communication channels over which the data flow from one vertex to another.

Moving computation to data is another advantage of the MapReduce and Dryad have over the other parallel programming runtimes. With the ever-increasing requirement of processing large volumes of data, we believe that this approach has a greater impact on the usability of the parallel programming runtimes in the future.

### 2.1. Current MapReduce Implementations

Google's MapReduce implementation is coupled with a distributed file system named Google File System (GFS) [22] where it reads the data for MapReduce computations and stores the results. According to the seminal paper by J. Dean et al.[3], in their MapReduce implementation, the intermediate data are first written to the local files and then accessed by the reduce tasks. The same architecture is adopted by the Apache's MapReduce implementation – Hadoop.

Hadoop stores the intermediate results of the computations in local disks, where the computation tasks are executed, and informs the appropriate workers to retrieve (pull) them for further processing. The same approach is adopted by Disco [23] – another open source MapReduce runtime developed using a functional programming language named Erlang [24]. Although this strategy of writing intermediate result to the file system makes the above runtimes robust, it introduces an additional step and a considerable communication overhead to the MapReduce computation, which could be a limiting factor for some MapReduce computations. Apart from the above, all these runtimes focus mainly on computations that utilize a single *map/reduce* computational unit. Iterative MapReduce computations are not well supported.

### 2.2. CGL-MapReduce

CGL-MapReduce is a novel MapReduce runtime that uses streaming for all the communications, which eliminates the overheads associated with communicating via a file system. The use of streaming enables the CGL-MapReduce to send the intermediate results directly from its producers to its consumers.

Currently, we have not integrated a distributed file system such as HDFS with CGL-MapReduce, but it can read data from a typical distributed file system such as NFS or from local disks of compute nodes of a cluster with the help of a meta-data file. The fault tolerance support for the CGL-MapReduce will harness the reliable delivery

mechanisms of the content dissemination network that we use. Figure 4 shows the main components of the CGL-MapReduce.

The CGL MapReduce runtime system is comprised of a set of workers, which perform *map* and *reduce* tasks and a content dissemination network that handles all the

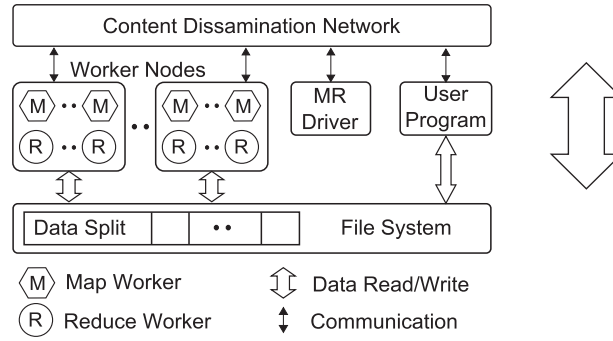


Figure 4: Components of the CGL-MapReduce System

underlying communications. As in other MapReduce runtimes, a master worker (*MRDriver*) controls the other workers according to instructions given by the user program. However, unlike typical MapReduce runtimes, CGL-MapReduce supports both single-step and iterative MapReduce computations.

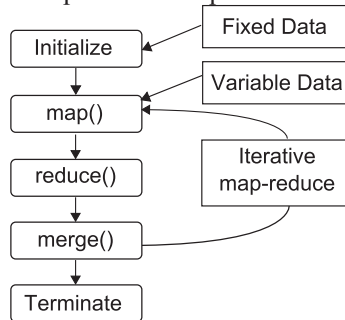


Figure 5. Computation phases of CGL-MapReduce

A MapReduce computation under CGL-MapReduce passes through several phases of computations as shown in figure 5. In CGL-MapReduce the initialization phase is used to configure both the *map/reduce* tasks and can be used to load any fixed data necessary for the *map/reduce* tasks. The *map* and *reduce* stages perform the necessary data processing while the framework directly transfers the intermediate result from *map* tasks to the *reduce* tasks. The *merge* phase is another form of reduction which is used to collect the results of the *reduce* stage to a single value. The User Program has access to the results of the *merge* operation. In the case of iterative MapReduce computations, the user program can call for another iteration of MapReduce by looking at the result of the merge operation and the framework performs another iteration of MapReduce using the already configured *map/reduce* tasks eliminating the necessity of configuring *map/reduce* tasks again and again as it is done in Hadoop.

CGL-MapReduce is implemented in Java and utilizes NaradaBrokering[25], a streaming-based content dissemination network. The CGL-MapReduce research prototype provides the runtime capabilities of executing MapReduce computations



written in the Java language. MapReduce tasks written in other programming languages require wrapper *map* and *reduce* tasks in order for them to be executed using CGL-MapReduce.

### 2.3. Performance Evaluation

To evaluate the different runtimes for their performance we have selected several data analysis applications. First, we applied the MapReduce technique to parallelize a High Energy Physics (HEP) data analysis application and implemented it using Hadoop, CGL-MapReduce, and Dryad (Note: The academic release of Dryad only exposes the DryadLINQ [26] API for programmers. Therefore, all our implementations are written using DryadLINQ although the underlying runtime it uses is Dryad). The HEP data analysis application processes large volumes of data and performs a histogramming operation on a collection of event files produced by HEP experiments. Next, we applied the MapReduce technique to parallelize a Kmeans clustering [27] algorithm and implemented it using Hadoop, CGL-MapReduce, and Dryad. Details of these applications and the challenges we faced in implementing them can be found in [28]. In addition, we implemented the same Kmeans algorithm using MPI (C++) as well. We have also implemented a matrix multiplication algorithm using Hadoop and CGL-MapReduce. We also implemented two common text-processing applications, which perform a “word histogramming” operation, and a “distributed grep” operation using Dryad, Hadoop, and CGL-MapReduce. Table 1 and Table 2 highlight the details of the hardware and software configurations and the various test configurations that we used for our evaluations.

**Table 2.** Test configurations.

Feature	HEP Data Analysis	Kmeans clustering	Matrix Multiplication	Histogramming & Grep
Cluster Ref	H	G	G	B
Number of Nodes	12	4	5	4
Number of Cores	96	32	40	32
Amount of Data	Up to 1TB of HEP data	Up to 10 million data points	Up to 16000 rows and columns	100GB of text data
Data Location	IU Data Capacitor: a high-speed and high-bandwidth storage system running the Lustre File System	Hadoop : HDFS CGL-MapReduce : NFS Dryad : Local Disc	Hadoop : HDFS CGL-MapReduce : NFS	Hadoop : HDFS CGL-MapReduce: Local Disc Dryad : Local Disc
Language	Java, C++ (ROOT)	Java, C++	Java	Java, C#

For the HEP data analysis, we measured the total execution time it takes to process the data under different implementations by increasing the amount of data. Figure 6 (a) depicts our results.

Hadoop and CGL-MapReduce both show similar performance. The amount of data accessed in each analysis is extremely large and hence the performance is limited by the I/O bandwidth of a given node rather than the total processor cores. The overhead induced by the MapReduce implementations has negligible effect on the overall computation.



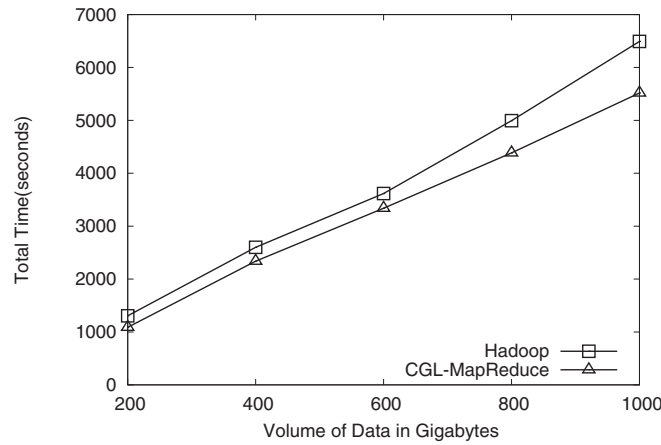


Figure 6(a). HEP data analysis, execution time vs. the volume of data (fixed compute resources)

The Dryad cluster (Table 1 ref. B) we used has a smaller hard disks compared to the other clusters we use. Therefore, to compare the performance of Hadoop, CGL-MapReduce, and Dryad for HEP data analysis, we have performed another test using a smaller data set on a smaller cluster configuration. Since Dryad is deployed on a Windows cluster running HPC Server Operating System(OS) while Hadoop and CGL-MapReduce are run on Linux clusters, we normalized the results of this benchmark to eliminate the differences caused by the hardware and the different OSs. Figure 6(b) shows our results.

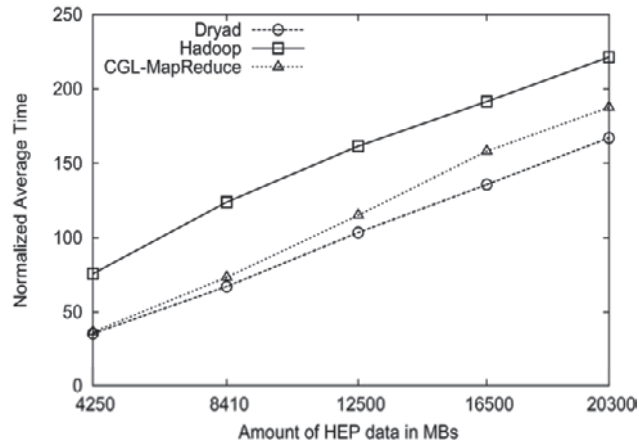


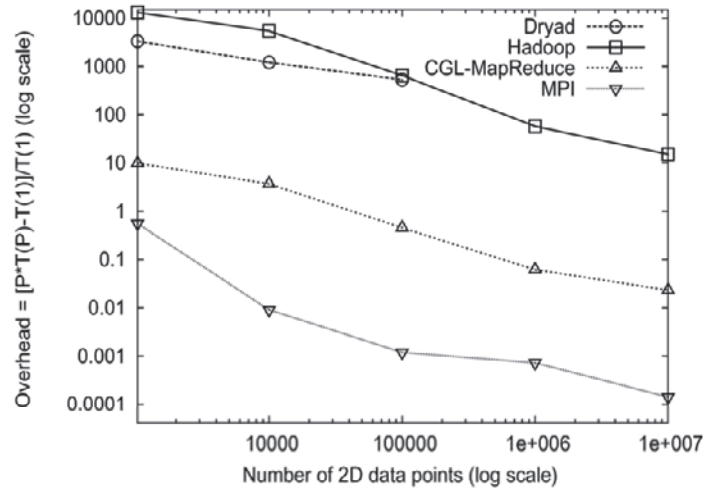
Figure 6(b). HEP data analysis, execution time vs. the volume of data (fixed compute resources). Note: In the Dryad version of HEP data analysis the “reduction” phase (combining of partial histograms produced by the “map” tasks) is performed by the GUI using a separate thread. So the timing results for Dryad does not contain the time for combining partial histograms.

Figure 6(a) and 6(b) show that Hadoop, Dryad, and CGL-MapReduce all perform nearly equally for the HEP data analysis. HEP data analysis is both compute and data intensive and hence the overheads associated with different parallel runtimes have negligible effect on the overall performance of the data analysis.

We evaluate the performance of different implementations for the Kmeans clustering application and calculated the parallel overhead ( $\phi$ ) induced by the different parallel programming runtime using the formula given below. In this formula  $P$  denotes the number of hardware processing units (i.e. number of cores used) and  $T(P)$  denotes the total execution time of the program when  $P$  processing units are used.  $T(1)$  denotes the total execution time for a single threaded program. Note  $\phi$  is just  $(1/\text{efficiency} - 1)$  and often is preferable to efficiency as overheads are summed linearly in  $\phi$ .

$$\phi(P) = [PT(P) - T(1)] / T(1) \quad (2.1)$$

Figure 7 depicts our performance results for Kmeans expressed as overhead.



**Figure 7.** Overheads associated with Hadoop, Dryad, CGL-MapReduce, and MPI for Kmeans clustering – iterative MapReduce - (Both axes are in log scale)

The results in figure 7 show that although the overheads of different parallel runtimes reduce with the increase in the number of data points, both Hadoop and Dryad have very large overheads for the Kmeans clustering application compared to MPI and CGL-MapReduce implementations.

Matrix multiplication is another iterative algorithm that we have implemented using Hadoop and CGL-MapReduce. To implement matrix multiplication using MapReduce model, we adopted the row/column decomposition approach to split the matrices. To clarify our algorithm let's consider an example where two input matrices  $A$  and  $B$  produce matrix  $C$  as the result of the multiplication process. We split the matrix  $B$  into  $n$  column blocks where  $n$  is equal to the number of map tasks used for the computation. The matrix  $A$  is split to  $m$  row blocks where  $m$  determines the number of iterations of MapReduce computations needed to perform the entire matrix multiplication. In each iteration, all the map tasks consume two inputs; (i) a column block of matrix  $B$  and (ii) a row block of matrix  $A$  and collectively they produce a row block of the resultant matrix  $C$ . The column block associated with a particular map task is fixed throughout the computation while the row blocks are changed in each iteration. However, in Hadoop's programming model, there is no way to specify this behavior and hence it

loads both the column block and the row block in each iteration of the computation. CGL-MapReduce supports the notion of long running map/reduce tasks where these task are allowed to retain static data in memory across invocations yielding better performance characteristics for iterative MapReduce computations.

For the matrix multiplication program, we measured the total execution time by increasing the size of the matrices used for the multiplication, using both Hadoop and CGL-MapReduce implementations. The result of this evaluation is shown in figure 8.

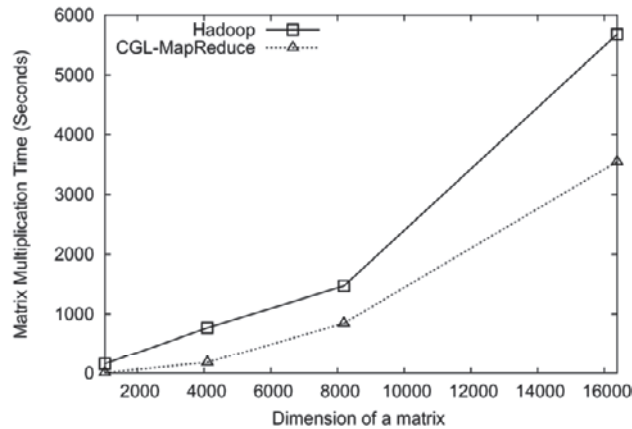


Figure 8. Performance of the Hadoop and CGL-MapReduce for matrix multiplication

The results in figure 7 and figure 8 show how the approach of configuring once and re-using of *map/reduce* tasks across iterations and the use of streaming have improved the performance of CGL-MapReduce for iterative MapReduce tasks. The communication overhead and the loading of static data in each iteration have resulted large overheads in iterative MapReduce computations implemented using Hadoop. The DAG based execution model of Dryad requires generation of execution graphs with fixed number of iterations. It also supports “loop unrolling” where a fixed number of iterations are performed as a single execution graph (a single query of DryadLINQ). The number of loops that can be unrolled is limited by the amount of stack space available for a process, which executes a collection of graph vertices as a single operation. Therefore, an application, which requires  $n$  iterations of MapReduce computations, can perform it in  $m$  cycles where in each cycle; Dryad executes a computation graph with  $n/m$  iterations. In each cycle the result computed so far is written to the disk and loaded back at the next cycle. Our results show that even with this approach there are considerable overheads for iterative computations implemented using Dryad.

The performance results of the two text processing applications comparing Hadoop, CGL-MapReduce, and Dryad are shown in figure 9 and figure 10.

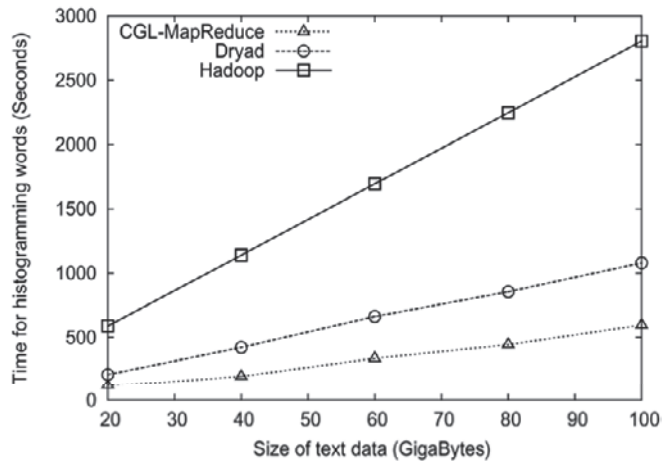


Figure 9. Performance of Dryad, Hadoop, and CGL-MapReduce for “histogramming of words” operation.

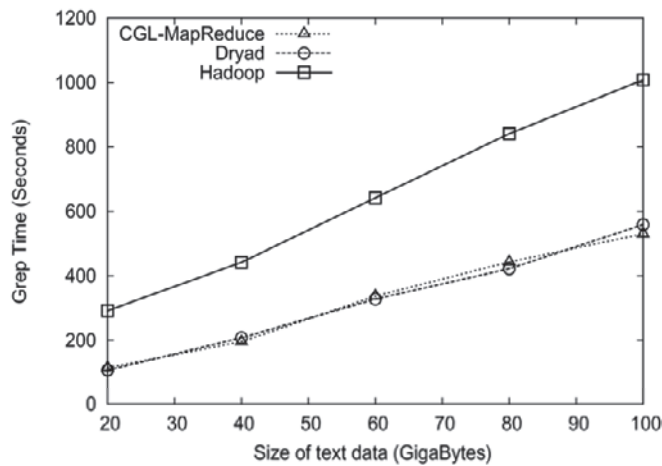
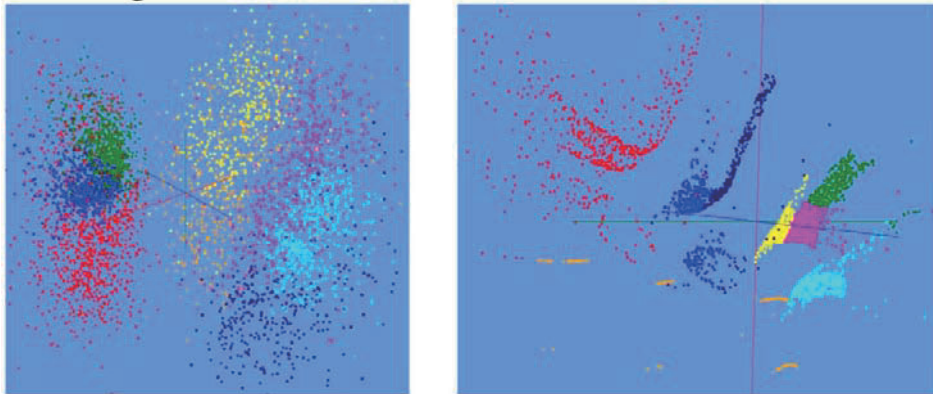


Figure 10. Performance of Dryad, Hadoop, and CGL-MapReduce for “distributed grep” operation

In both these tests, Hadoop shows higher overall processing time compared to Dryad and CGL-MapReduce. This could be mainly due to its distributed file system and the file based communication mechanism. Dryad uses in memory data transfer for intra-node data transfers and a file based communication mechanism for inter-node data transfers whereas in CGL-MapReduce all data transfers occur via streaming. The “word histogramming” operation requires higher data transfer requirements compared to the “distributed grep” operation and hence the streaming data transfer approach adopted by the CGL-MapReduce shows lowest execution times for the “word histogramming” operation. In “distributed grep” operation both Dryad and CGL-MapReduce show close performance results.

### 3. Multidimensional Scaling MDS

Dimension reduction algorithms are used to reduce dimensionality of high dimensional data into Euclidean low dimensional space, so that dimension reduction algorithms are used as visualization tools. Some dimension reduction approaches, such as generative topographic mapping (GTM) [29] and Self-Organizing Map (SOM) [30], seek to preserve topological properties of given data rather than proximity information. On the other hand, multidimensional scaling (MDS) [31-32] tries to maintain dissimilarity information between mapping points as much as possible. The MDS algorithm involves several full  $N \times N$  matrices where we are mapping  $N$  data points. Thus, the matrices could be very large for large problems ( $N$  could be as big as millions even today). For large problems, we will initially cluster the given data and use the cluster centers to reduce the problem size. Here we parallelize an elegant algorithm for computing MDS solution, named SMACOF (Scaling by MAjorizing a COMplicated Function) [33-34], using MPI.NET [35-36] which is an implementation of message passing interface (MPI) for C# language and presents performance analysis of the parallel implementation of SMACOF on multicore cluster systems. We show some examples of the use of MDS to visualize the results of the clustering algorithms of section 4 in figure 11. These are datasets in high dimension (from 20 in figure 11(right) to over a thousand in figure 11(left)) which are projected to 3D using proximity (distance/dissimilarity) information. The figure shows 2D projections determined by us from rotating 3D MDS results.



**Figure 11.** Visualization of MDS projections using parallel SMACOF described in section 3. Each color represents a cluster determined by the PWDA algorithm of section 4. Figure 11(left) corresponds to 4500 ALU pairwise aligned Gene Sequences with 8 clusters [37] and 11(right) to 4000 Patient Records with 8 clusters from [38]

Multidimensional scaling (MDS) is a general term for a collection of techniques to configure data points with proximity information, typically dissimilarity (interpoint distance), into a target space which is normally Euclidean low-dimensional space. Formally, the  $N \times N$  dissimilarity matrix  $\Delta = (\delta_{ij})$  should be satisfied symmetric ( $\delta_{ij} = \delta_{ji}$ ), nonnegative ( $\delta_{ij} \geq 0$ ), and zero diagonal elements ( $\delta_{ii} = 0$ ) conditions. From given dissimilarity matrix  $\Delta$ , a configuration of points is constructed by the MDS algorithm in a Euclidean target space with dimension  $p$ . The output of MDS algorithm can be an  $N \times p$  configuration matrix  $X$ , whose rows represent each data point  $x_i$  in Euclidean  $p$ -dimensional space. From configuration matrix  $X$ , it is easy to compute the Euclidean interpoint distance  $d_{ij}(X) = \|x_i - x_j\|$  among  $N$  configured points in the target space and

to build the  $N \times N$  Euclidean interpoint distance matrix  $D(X) = (d_{ij}(X))$ . The purpose of MDS algorithm is to map the given points into the target  $p$ -dimensional space, while the interpoint distance  $d_{ij}(X)$  is approximated to  $\delta_{ij}$  with different MDS forms correspondingly to different measures of the discrepancy between  $d_{ij}(X)$  and  $\delta_{ij}$ . STRESS [39] and SSTRESS [40] were suggested as objective functions of MDS algorithms. STRESS ( $\sigma$  or  $\sigma(X)$ ) criterion (Eq. (3.1)) is a weighted squared error between distance of configured points and corresponding dissimilarity, but SSTRESS ( $\sigma^2$  or  $\sigma^2(X)$ ) criterion (Eq. (3.2)) is a weighted squared error between squared distance of configured points and corresponding squared dissimilarity.

$$\sigma(X) = \sum_{i < j \leq n} w_{ij} (d_{ij}(X) - \delta_{ij})^2 \quad (3.1)$$

$$\sigma^2(X) = \sum_{i < j \leq n} w_{ij} [(d_{ij}(X))^2 - (\delta_{ij})^2]^2 \quad (3.2)$$

where  $w_{ij}$  is a weight value, so  $w_{ij} \geq 0$ . Therefore, the MDS can be thought of as an optimization problem, which is minimization of the STRESS or SSTRESS criteria during constructing a configuration of points in the  $p$ -dimension target space.

### 3.1. Scaling by MAjorizing a COmplicated Function (SMACOF)

Scaling by MAjorizing a COmplicated Function (SMACOF) [33-34] is an iterative majorization algorithm in order to minimize objective function of MDS. SMACOF is likely to find a local not global minima as is well known from gradient descent methods. Nevertheless, it is powerful since it guarantees a monotonic decrease of the objective function. The procedure of SMACOF is described in Algorithm 1. For the mathematical details of SMACOF, please refer to [32].

---

#### Algorithm 1 SMACOF algorithm

---

```

 $Z \leftarrow X^{[0]}$ ;
 $k \leftarrow 0$ ;
 $\varepsilon \leftarrow$  small positive number;
 $MAX \leftarrow$  maximum iteration;
Compute  $\sigma^{[0]} = \sigma(X^{[0]})$ ;
while  $k = 0$  or  $(\Delta\sigma(X^{[k]})) > \varepsilon$  and  $k \leq MAX$  do
   $k \leftarrow k + 1$ ;
   $X^{[k]} = V^\dagger B(X^{[k-1]})X^{[k-1]}$ 
  Compute  $\sigma^{[k]} = \sigma(X^{[k]})$ 
   $Z \leftarrow X^{[k]}$ ;
end while
return  $Z$ ;

```

---

### 3.2. Distributed-Memory Parallel SMACOF

In order to implement distributed-memory parallel SMACOF, one must address two issues: one is the data decomposition where we choose block matrix decomposition for our SMACOF implementation since it involves matrix multiplication iterated over

successive gradient descents, and the other is the required communication between decomposed processes. For the data decomposition, our implementation allows users to choose the number of row-blocks and column-blocks with a constraint that the product of the number of row-blocks and column-blocks should be equal to the number of processes, so that each process will be assigned corresponding decomposed sub-matrix. For instance, if we run this program with 16 processes, then users can decompose the  $N \times N$  full matrices into not only  $4 \times 4$  block matrices but also  $16 \times 1$ ,  $8 \times 2$ ,  $2 \times 8$ , and  $1 \times 16$  block matrices. In addition, message passing interface (MPI) is used to communicate between processes, and MPI.NET is used for the communication.

### 3.2.1. Advantages of Distributed-memory Parallel SMACOF

The running time of SMACOF algorithm is  $O(N^2)$ . Though matrix multiplication of  $V^T \cdot B(X)$  takes  $O(N^3)$ , you can reduce the computation time by using associativity of matrix multiplication. By the associative property of the matrix multiplication,  $(V^T \cdot B(X)) \cdot X$  is equal to  $V^T \cdot (B(X) \cdot X)$ . While the former takes the order of  $O(N^3 + N^2p)$ , the latter takes only  $O(2N^2p)$ , where  $N$  is the number of points and  $p$  is the target dimension that we would like to find a configuration for given data. Normally, the target dimension  $p$  is two or three for the visualization, so  $p$  could be considered as a constant for computational complexity. Also, SMACOF algorithm uses at least four full  $N \times N$  double matrices, i.e.  $A$ ,  $D$ ,  $V^T$ , and  $B(X)$ , which means at least  $32 \times N^2$  bytes of memory should be allocated to run SMACOF program.

As in general, there are temporal and spatial advantages when we use distributed-memory parallelism. First, computational advantage should be achieved by both shared-memory and distributed-memory parallel implementation of SMACOF. While shared-memory parallelism is limited by the number of processors (or cores) in a single machine, distributed-memory parallelism can be extended the available number of processors (or cores) as much as machines are available, theoretically. SMACOF algorithm uses at least  $32 \times N^2$  bytes of memory as we mentioned above. For example, 32MB, 3.2GB, 12.8GB, and 320GB are necessary for  $N = 1000$ , 10000, 20000, 100000, correspondingly. Therefore, a multicore workstation, which has a 8GB of memory will be able to run SMACOF algorithm with 10000 data points. However, this workstation cannot be used to run the same algorithm with 20000 data points. Shared memory parallelism increases performance but does not increase size of problem that can be addressed. Thus, the distributed-memory parallelism allows us to run SMACOF algorithm with much more data, and this benefit is quite important in the era of a data deluge.

### 3.3. Experimental Results and Analysis

For the performance experiments of the distributed-memory parallel SMACOF, we use two nodes of Ref C and one node of Ref D in Table 1. For the performance test, we generate artificial random data set which is in 8-centered Gaussian distribution in 4-dimension with different number of data points, such as 128, 256, 512, 1024, 2048, and 4096.

Due to gradient descent attribute of SMACOF algorithm, the final solution highly depends on the initial mapping. Thus, it is appropriate to use random initial mapping for the SMACOF algorithm unless specific prior initial mapping exists, and to run several times to increase the probability to get better solution. If the initial mapping is



different, however, the computation amount can be varied whenever the application runs, so that we could not measure any performance comparison between two experimental setups, since it could be inconsistent. Therefore, the random seed is fixed for the performance measures of this paper to generate the same answer and the same necessary computation for the same problem. The stop condition threshold value ( $\epsilon$ ) is also fixed for each data. We will investigate the dependence on starting point more thoroughly using other approaches discussed in section 3.4.

### 3.3.1. Performance Analysis

For the purpose of performance comparison, we implemented the sequential version of SMACOF algorithm. The sequential SMACOF is executed on each test node, and the test results are in Table 3. Note that the running time of **D** is almost twice faster than the other two nodes, though the core's clock speed of each node is similar. The reason would be the cache memory size. L2 cache of two Ref C nodes (**C1** and **C2**) is much smaller than that of **D** node.

**Table 3.** Sequential Running time in seconds on each test node

Data size	C1	C2	D
128	0.3437	0.3344	0.1685
256	1.9031	1.9156	0.9204
512	9.128	9.2312	4.8456
1024	32.2871	32.356	18.1281
2048	150.5793	150.949	83.4924
4096	722.3845	722.9172	384.7344

Initially we measured the performance of the distributed-memory parallel SMACOF (MPI\_SMACOF) on each test node only. Figure 12 shows the speedup of each test node with different number of processes. Both axes of the Figure 12 are in logarithmic scale. As the Figure 12 depicted, the MPI\_SMACOF is not good for small data, such as 128 and 256 data points. However, for larger data, i.e. 512 and more data points, the MPI\_SMACOF shows great performance on the test data. You should notice those speedup values of larger data, such as 1024 or more data points on **C1** and **C2** nodes are bigger than the actual processes number using the MPI\_SMACOF application, which corresponds to super-linear speedup. However, on the **D** node, it represented good speedup but not super-linear speedup at all. The reason of super-linear speedup is related to cache-hit ratio, as we discussed about sequential running results. MPI\_SMACOF implemented in the way of block decomposition, so that those sub-matrix would be better matched in the cache line size and the portion of sub-matrix which is in cache memory at a moment would be bigger than the portion of whole matrix in it. The Figure 12 also describes that the speedup ratio (or efficiency) becomes worse when you run MPI\_SMACOF with more processes on single node. It seems natural that as the number of computing units increases, the assigned computing job will be decreased but the communication overhead will be increased.

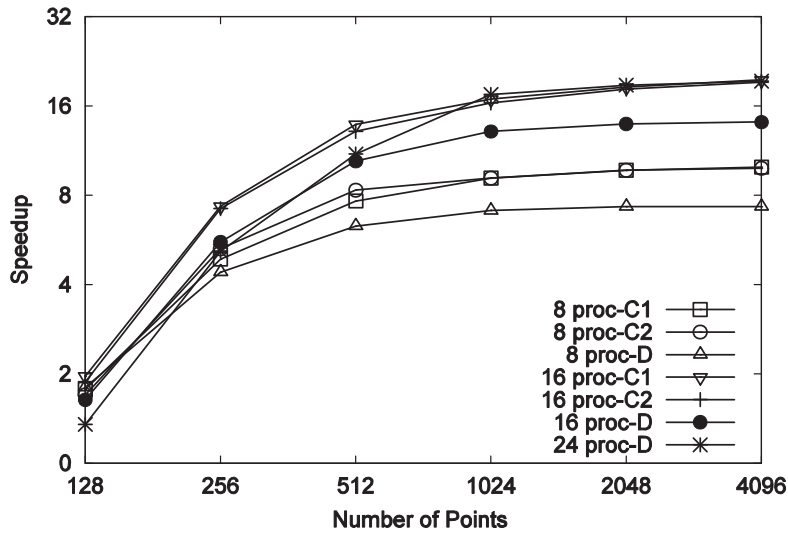


Figure 12. Speedup of MPI\_SMACOF performance on each test node

In addition, we have measured the performance of the proposed MPI\_SMACOF algorithm on all the three test nodes with different number of processes. Figure 13 illustrates the speedup of those experiments with respect to the average of the sequential SMACOF running time on each node. The comparison with average might be reasonable since, for every test case, the processes are equally spread as much as possible on those three test nodes except the case of 56 processes running. The Figure 13 represents that the speedup values are increasing as the data size is getting bigger. This result shows that the communication overhead on different nodes is larger than communication overhead on single node, so that the speedup is still increasing, even with large test data such as 2048 and 4096 points, instead of being converged as in Figure 12.

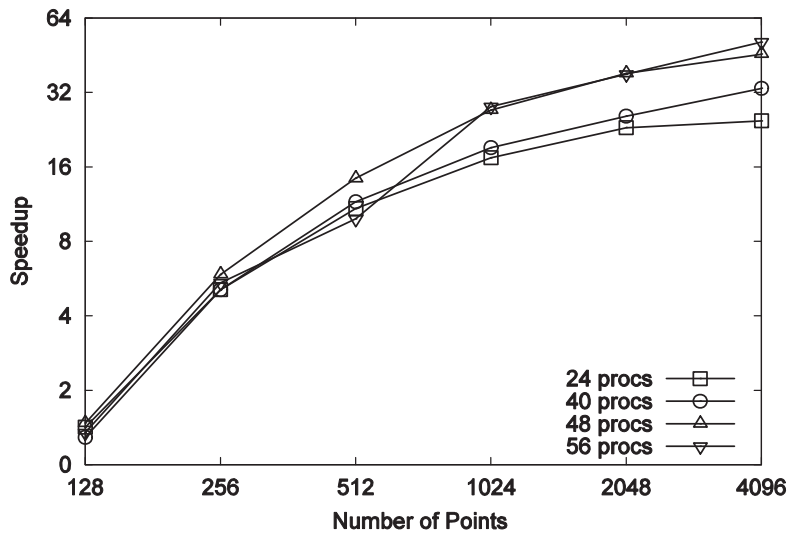


Figure 13. Speedup of MPI\_SMACOF on combine nodes

### 3.4. Conclusions

We have developed a dimension mapping tool that is broadly applicable as it only uses dissimilarity values and does not require the points to be in a vector space. We have good parallel performance and are starting to use it for science applications as illustrated in figure 11. In later work, we will compare the method described with alternatives that can also be parallelized and avoid the steepest descent approach of SMACOF which can lead to local minima. One approach, first described in [41] and [42], uses deterministic annealing based on ideas sketched in section 4. This still uses Expectation Maximization (EM) (steepest descent) but only for the small steps needed as temperature is decreased. We will also implement the straightforward but possibly best method from ref [43] that solves equations (3.1) and (3.2) as  $\chi^2$  problems and uses optimal solution methods for this.

## 4. Multicore Clustering

### 4.1. Algorithms

Clustering can be viewed as an optimization problem that determines a set of K clusters by minimizing

$$H_{\text{VECD A}} = \sum_{i=1}^N \sum_{k=1}^K M_i(k) D_{\text{VEC}}(i,k) \quad (4.1)$$

where  $D_{\text{VEC}}(i,k)$  is the distance between point  $i$  and cluster center  $k$ .  $N$  is the number of points and  $M_i(k)$  is the probability that point  $i$  belongs to cluster  $k$ . This is the vector version and one obtains the pairwise distance model with:

$$H_{\text{PWDA}} = 0.5 \sum_{i=1}^N \sum_{j=1}^N D(i,j) \sum_{k=1}^K M_i(k) M_j(k) / C(k) \quad (4.2)$$

and  $C(k) = \sum_{i=1}^N M_i(k)$  is the expected number of points in the  $k$ 'th cluster.  $D(i,j)$  is pairwise distance between points  $i$  and  $j$ . Equation (4.1) requires one be able to calculate the distance between a point  $i$  and the cluster center  $k$  and this is only possible when one knows the vectors corresponding to the points  $i$ . (4.2) reduces to (4.1) when one inserts vector formulae and drops terms that average to zero. The formulation (4.2) is important as there are many important clustering applications where one only knows distances between points and not a Euclidean vector representation.

One must minimize (4.1) or (4.2) as a function of cluster centers for case VECD A and cluster assignments  $M_i(k)$  for case PWDA. One can derive deterministic annealing from an informatics theoretic [17] or physics formalism [18]. In latter case one smoothes out the cost function (4.1) or (4.2) by averaging with the Gibbs distribution  $\exp(-H/T)$ . This implies in a physics language that one is minimizing not  $H$  but the free energy  $F$  at temperature  $T$  and entropy  $S$

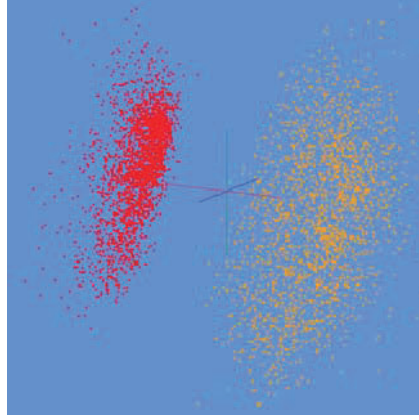
$$F = H - TS \quad (4.3)$$

For VECD A and Hamiltonian  $H$  given by equation (4.1), one can do this averaging exactly.

$$M_i(k) = \exp(-D_{\text{VEC}}(i,k)/T) / Z_i \quad (4.4)$$

$$Z_i = \sum_k \exp(-D_{\text{VEC}}(i,k)/T) \quad (4.5)$$

$$F = -T \sum_{i=1}^N \log [Z_i] / N \quad (4.6)$$



**Figure 14.** Preliminary stage of clustering shown in figure 11(left) corresponding to 4500 ALU pairwise aligned Gene Sequences with 2 clusters [37]

For the case of equation (4.2) where only distances are known, the integrals with the Gibbs function are intractable analytically as the degrees of freedom  $M_i(k)$  appear quadratically in the exponential. In the more familiar simulated annealing approach to optimization, these integrals are effectively performed by Monte Carlo. This implies simulated annealing is always applicable but is usually very slow. The applicability of deterministic annealing was enhanced by the important observation in [18] that one can use an approximate Hamiltonian  $H_0$  and average with  $\exp(-H_0/T)$ . For pairwise clustering (4.2), one uses the form motivated by the VECDA formalism (4.4).

$$H_0 = \sum_{i=1}^N \sum_{k=1}^K M_i(k) \varepsilon_i(k) \quad (4.7)$$

$$M_i(k) \propto \exp(-\varepsilon_i(k)/T) \text{ with } \sum_{k=1}^K M_i(k) = 1 \quad (4.8)$$

$\varepsilon_i(k)$  are new degrees of freedom. This averaging removes local minima and is designed so that at high temperatures one starts with one cluster. As temperature is lowered one minimizes the Free Energy (4.3) with respect to the degrees of freedom. A critical observation of Rose [17] allows one to determine when to introduce new clusters. As in usual expectation maximization (steepest descent) the first derivative of equation (4.3) is set to zero to find new estimates for  $M_i(k)$  and other parameters such as cluster centers for VECDA. Then one looks at the second derivative  $\Gamma$  of  $F$  to find instabilities that are resolved by splitting clusters. One does not examine the full matrix but the submatrices coming from restricting  $\Gamma$  to variations of the parameters of a single cluster with the  $K-1$  other clusters fixed and multiple identical clusters placed at location of clusters whose stability one investigates. As temperature is lowered one finds that clusters naturally split and one can easily understand this from the analytic form for  $\Gamma$ . The previous work [18] on PWDA was incomplete and did not consider calculation of  $\Gamma$  but rather only assumed an a priori fixed number of clusters. We have completed the formalism and implemented it in parallel. Note we only need to find the single lowest eigenvalue of  $\Gamma$  (restricted to varying one cluster). This is implemented as power (Arnoldi) method. One splits the cluster if its restricted  $\Gamma$  has a negative eigenvalue and this is the smallest when looked at over all clusters.

The formalism for VECDA can be found in our earlier work and [17]. Here we just give results for the more complex PWDA and use it to illustrate both methods. We let indices  $k \mu \lambda$  runs over clusters from 1 to  $K$  while  $i j \alpha \beta$  run over data points from 1 to  $N$ .  $M_i(k)$  has already been given in equation (4.8). Then one calculates:

$$A(k) = -0.5 \sum_{i=1}^N \sum_{j=1}^N D(i, j) M_i(k) M_j(k) / C(k)^2 \quad (4.9a)$$

$$B_\alpha(k) = \sum_{i=1}^N D(i, \alpha) M_i(k) / C(k) \quad (4.9b)$$

$$C(k) = \sum_{i=1}^N M_i(k) \quad (4.9c)$$

$$\text{Allowing one to derive the estimate } \varepsilon_\alpha(k) = (B_\alpha(k) + A(k)) \quad (4.10)$$

Equation (4.10) minimizes  $F$  of equation (4.3). The  $NK \times NK$  second derivative matrix  $\Gamma$  is given by:

$$\begin{aligned} \{\alpha, \mu\} \Gamma_{\{\beta, \lambda\}} = & (1/T) \delta_{\alpha\beta} \{M_\alpha(\mu) \delta_{\mu\lambda} - M_\alpha(\mu) M_\alpha(\lambda)\} + (M_\alpha(\mu) M_\beta(\lambda) / T^2) \{ \sum_{k=1}^K [ \\ & -2A(k) - B_\beta(k) - B_\alpha(k) + D(\alpha, \beta) ] [M_\alpha(k) - \delta_{k\mu}] [M_\beta(k) - \delta_{k\lambda}] / C(k) \} \end{aligned} \quad (4.11)$$

Equations (4.9) and (4.10) followed by (4.8) represent the basic steepest descent iteration (Expectation Maximization) that is performed at fixed temperature until the estimate for  $\varepsilon_\alpha(k)$  is converged. Note steepest descent is a reasonable approach for deterministic annealing as one has smoothed the cost function to remove (some) local minima. Then one decides whether to split a cluster from the eigenvalues of  $\Gamma$  as discussed above. If splitting is not called for, one reduces the temperature and repeats equations (4.8) through (4.11). There is an elegant method of deciding when to stop based on the fractional freezing factors  $\Phi(k)$

$$\Phi(k) = \sum_{i=1}^N M_i(k) (1 - M_i(k)) / C(k) \quad (4.12)$$

As temperatures are lowered after final split, then the  $M_i(k)$  tend to either 0 or 1 so  $\Phi(k)$  tends to zero. We currently stop when all the freezing factors are  $< 0.002$  but obviously this precise value is ad-hoc.

#### 4.2. Multi-Scale and Deterministic Annealing

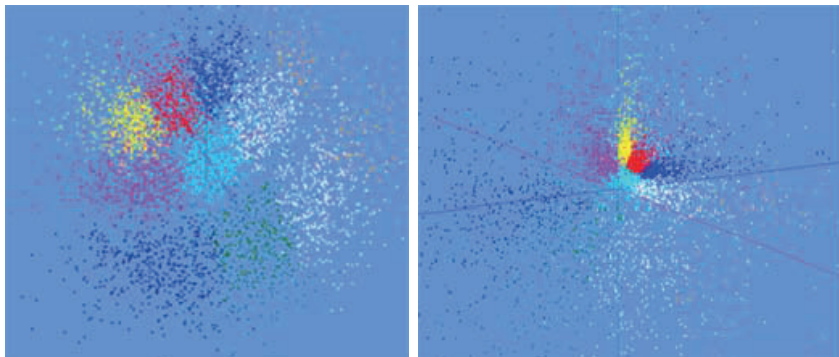
In references [12] and [14], we explain how a single formalism describes many different problems: VECDA (Clustering of points defined by vectors with deterministic annealing) [16-17], Gaussian Mixture Models (GMM) [44]; Gaussian Mixture Models with deterministic annealing (GMMDA) [45]; and Generative Topographic Maps (GTM) [29]. One can also add deterministic annealing to GTM and we are currently working on this for Web applications [46]. Deterministic annealing can be considered as a multi-scale approach as quantities are weighted by  $\exp(-D/T)$  for distances  $D$  and temperature  $T$ . Thus at a given temperature  $T$ , the algorithm is only sensitive to distances  $D$  larger than or of order  $T$ . One starts at high temperatures (determined by largest distance scale in problem) and reduce temperature (typically by 1% each iteration) until you reach either the distance scale or number of clusters desired. As explained in original papers [16], clusters emerge as phase transitions as one lowers the temperature and need not be put in by hand. For example the eight clusters in figure 11(left) were found systematically with clusters being added as one reduced temperature so that at a higher temperature one first split from one to two clusters to

find results of figure 14. The splits are determined from the structure of second derivative matrix equation (4.11) and figure 11(left) is for example found by continuing to reduce the temperature from intermediate result in figure 14.

#### 4.3. Operational Use of Clustering and MDS

The original data is clustered with VECDA (see earlier papers for examples) or PWDA and then visualized by mapping points to 3D with MDS as described in section 3 and visualizing with a 3D viewer written in DirectX. As a next step, we will allow users to select regions either from clustering or MDS and drill down into the substructure in this region. Like the simpler linear principal component analysis, MDS of a sub-region is generally totally different from that of full space. We note here that deterministic annealing can also be used to avoid local minima in MDS [47]. We will report our extensions of the original approach in [41-42] and comparison with Newton's method for MDS [43] elsewhere.

Clustering in high dimensions  $d$  is not intuitive geometrically as the volume of a cluster of radius  $R$  is proportional to  $R^{(d+1)}$  implying that a cluster occupying 0.1% of total volume has a radius reduced by only a factor 0.99 from that of overall space with  $d=1000$  (a value typical of gene sequences). These conceptual difficulties are avoided by the pairwise approach. One does see the original high dimension when projecting points to 3D for visualization as they tend to appear on surface of the lower dimensional space. This can be avoided as discussed in [42] by a mapping Distance  $D \rightarrow f(D)$  where  $f$  is a monotonic function designed so that the transformed distances  $f(D)$  are distributed uniformly in a lower  $d_L$  dimensional space. We experimented with  $d_L = 2$  and 4 where the mapping is analytically easy but found it did not improve the visualization. Typical results are shown in figure 15(right) that maps data of figure 15(left) to 2 dimensions before applying MDS – the clustering is still performed on original unmapped data. Certainly the tendency in figure 15(left) to be at edge of visualization volume is removed but data understanding does not seem improved. This approach finds an effective dimension  $d_{eff}$  for original data by comparing mean and standard deviation of all the inter-point distances  $D(i,j)$  with those in a dimension  $d_{eff}$ . This determines an effective dimension  $d_{eff}$  of 40-50 for sequence data and about 5 for medical record data; in each case  $d_{eff}$  is a dimension smaller than that of underlying vector space. This is not surprising as any data set is a very special correlated set of points.



**Figure 15:** Results of Clustering of 4500 ALU sequences into 10 clusters before (left) and after (right) dimensional reduction described in text below.



#### 4.4. Parallelism

The vector clustering model is suitable for low dimensional spaces such as our earlier work on census data [12] but the results of figures 11, 14 and 15 correspond to our implementation of PWDA – the pairwise distance clustering approach of [18] which starts from equation (4.2) and its structure has similarities to familiar  $O(N^2)$  problems such as (astrophysical) particle dynamics. As  $N$  is potentially of order a million we see that both MDS and pairwise clustering are potential supercomputing data analysis applications. The parallelism for clustering is straightforward data parallelism with the  $N$  points divided equally between the  $P$  parallel units. This is the basis of most MapReduce algorithms and clustering was proposed as a MapReduce application in [7]. We have in fact compared simple (K-means) clustering between versions and MapReduce and MPI in section 2 and ref. [28]. Note that VECDA should be more suitable than K-means for MapReduce as it has more computation at each iteration (MapReduce has greater overhead than MPI on communication and synchronization as shown in section 2). VECDA only uses reduction, barrier and broadcast operations in MPI and in fact MPI implementation of this algorithm is substantially simpler than the threaded version. Reduction, Barrier and Broadcast are all single statements in MPI but require several statements – especially for reduction – in the threaded case. Reduction is not difficult in threaded case but requires care with many opportunities for incorrect or inefficient implementations.

PWDA is also data parallel over points and its  $O(N^2)$  structure is tackled similarly to other  $O(N^2)$  algorithms by dividing the points between parallel units. Each MPI process also stores the distances  $D(i, j)$  for all points  $i$  for which process is responsible. Of course the threads inside this process can share all these distances stored in common memory of a multicore node. There are subtle algorithms familiar from N-body particle dynamics where a factor of 2 in storage (and in computation) is saved by using the symmetry  $D(i, j) = D(j, i)$  but this did not seem useful in this case. The MPI parallel algorithm now needs MPI\_SENDRECV to exchange information about the distributed vectors; i.e. one needs to know about all components of vectors  $M_i B_i$  and the vector  $A_i$  iterated in finding maximal eigenvectors. This exchange of information can either be done with a broadcast or as in results reported here by send-receive in ring structure as used in  $O(N^2)$  particle dynamics problems. We measured the separate times in the four components of MPI – namely send-receive, Reduction, and Broadcast and only the first two are significant reaching 5-25% of total time with Broadcast typically less than 0.1% of execution time. The time needed for MPI send-receive is typically 2 to 3 times that for reduction but the latter is a non trivial overhead (often 5-10%). Obviously broadcast time would go up if it was used in place of send-receive in information exchange step.

#### 4.5. Computational Complexity

The vector and pairwise clustering methods have very different and complementary computational complexities. VECDA execution time is proportional to  $N d^2$  for  $N$  points – each of dimension  $d$ . PWDA has an execution time proportional to  $N^2$ . PWDA can rapidly become a supercomputer computation. For example with 4500 sequence data points and 8 clusters, the sequential execution time is about 15 hours on a single core of the systems used in our benchmarks. A direct clustering with PWDA of half million points (relevant even today) would thus naturally use around 5000 cores



(100 points per core) with pure MPI parallelization. The hybrid threading-MPI parallelism could efficiently support more cores.

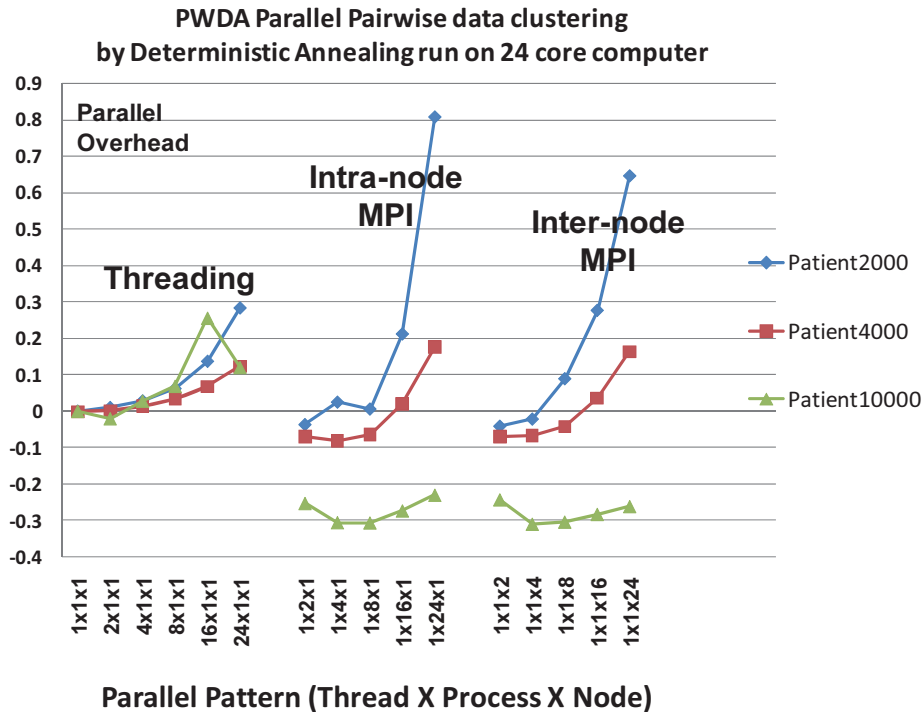
We note that currently some 40-70% of the computation time is used in deciding whether to split clusters in PWDA; there are probably significantly faster algorithms here. The runs of VECDA reported here correspond to a low dimension space  $d = 2$  for which negligible time is spent in splitting decision. The second derivative matrices are of size  $NK \times NK$  for PWDA and of size  $dK \times dK$  for VECDA. These are full matrices but as power method for determining maximal eigenvalues is used the computation is proportional to the square of the matrix dimension. For computations reported here, the annealing uses from 1000-10,000 temperature steps while each eigenvalue determination uses 10-200 iterations.

#### 4.6. Performance

We have performed extensive performance measurements [11-14] showing the effect of cache and for Windows runtime fluctuations can be quite significant. Here we give some typical results with figure 15 showing the performance of PWDA on the single 24 core workstation (ref D of table 1). The results are expressed as an overhead using the definitions of equation (1) introduced in section 2. We compare both MPI and thread based parallelism using Microsoft's CCR package [20-21]. As these codes are written in C#, we use MPI.NET[35-36] finding this to allow an elegant object-based extension of traditional MPI and good performance. MPI.NET is a wrapper for the production Microsoft MPI.

Figure 16 shows that although threading and MPI both get good performance, their systematics are different. For the extreme case of 24-way parallelism, the thread implementation shows an overhead that varies between 10 and 20% depending on the data set size. MPI shows a large overhead for small datasets that decreases with increasing dataset size so in fact 24-way MPI parallelism is 20% faster than the thread version on the largest 10,000 element dataset. This is due to the different sources of the overhead. For MPI the overhead is due to the communication calls which are due to reduce (20%) and send-receive (80%) and this as expected decreases (inversely proportional to dataset size) as the dataset size increases. For threads there is no memory movement overhead but rather the overhead is due to the Windows thread scheduling that leads to large fluctuations that can have severe effects on tightly synchronized parallel codes such as those in this paper as discussed in refs. [11-14]. We see some cases where the overhead is negative (super-linear speedup) which is due to better use of cache in the higher parallelism cases compared to sequential runs. This effect is seen in all our runs but differs between the AMD and Intel architectures reflecting their different cache size and architecture.

Comparing center and right datasets we see that MPI gets comparable performance on cores of a single node (center points) or when running one process per node on up to 24 nodes of the Infiniband connected cluster. In the results plotted in the figure. MPI gets better performance (smaller overhead) than threading on the largest 10,000 element Patient dataset. This reflects the large chunks of processing per MPI process. As seen in figure this is not always the case as threading outperforms MPI on the 2000 and 4000 element datasets for largest 24-way parallelism. As a dramatic example using all 768 cores of Tempest (ref I Table 1), the pattern 24X1X32 (24 threads on each of 32 nodes connected as 32 MPI processes) runs 172 times faster than the communication dominated 1X24X32 (24 internal MPI processes on each of 32 nodes).



**Figure 16.** Parallel Overhead for pure threading or pure MPI on Tempest (ref I of Table 1) for three different patient datasets with 2000, 4000 and 10,000 elements. The center and rightmost results are MPI.NET runs labeled 1XNX1 (center) or 1X1XN for N MPI processes. The leftmost results are CCR threading labeled NX1X1 for N threads. Left and center are run on one node; the right is one process per node on up to 24 nodes.

The fluctuations in thread execution times are illustrated in figure 17 showing standard deviations from 5 to 10% on a simple kernel representative of the VECDA clustering algorithm. The identical code (translated from C# to C) shows order of magnitude lower fluctuations when run under Linux [13] with interesting systematics even in Linux case. These fluctuations can give significant parallel overheads as parallel algorithms used in VECDA and PWDA like those in most scientific algorithms requires iterative thread synchronization at the rendezvous points. Here the execution time will be the maximum over that of all the simultaneous fluctuating threads and so increase as this number increases. As described in the earlier papers we have always seen this and reported this effect to Microsoft. We found that these fluctuations were the only sizeable new form of parallel overhead compared to those well known from traditional parallel computing i.e. in addition to load imbalance and communication overhead. We did note extra overheads due to different threads interfering on a single cache line (“false sharing”) but our current software is coded to avoid this.

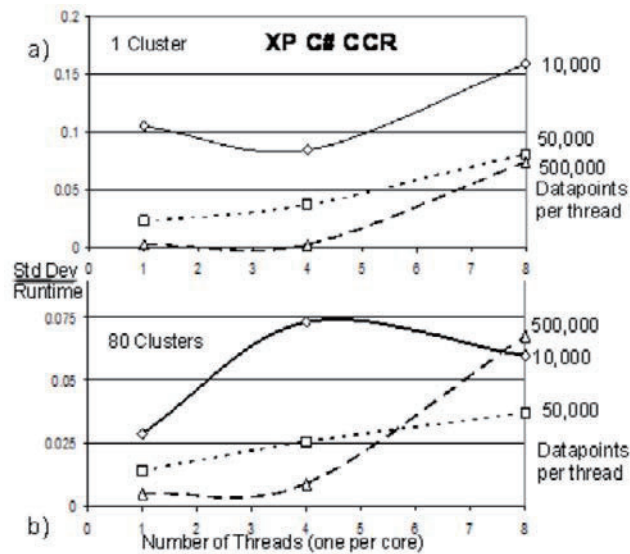


Figure 17. Measurements from [11, 12] showing 5 to 10% runtime fluctuations on an 8 core workstation. The results are plotted as a function of number of simultaneous threads from 1 to 8 and for three different dataset sizes.

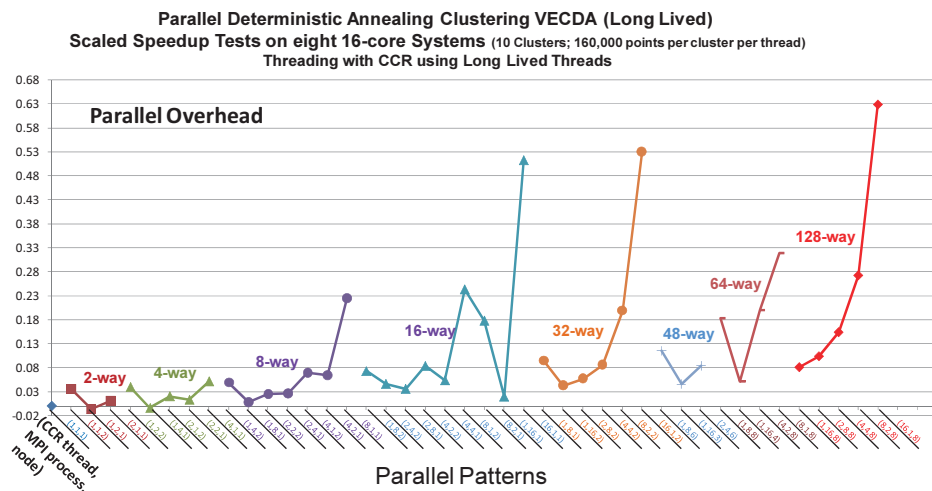
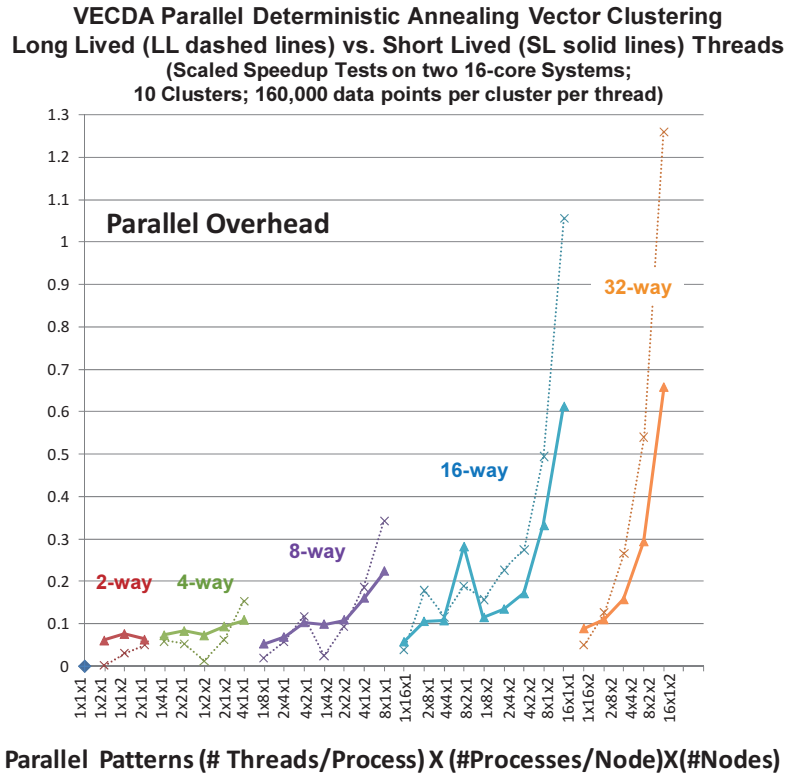


Figure 18. Parallel Overhead for VECDA using long lived threads run on 128 core Madrid Cluster in table 1. The results achieve a given parallelism by choosing number of nodes, MPI processes per node and threads per MPI process. The number of threads increases as you move from left to right for given level of parallelism.



**Figure 19.** Comparison of use of short lived (solid lines) and long lived (dashed lines) threads for the Vector-based deterministic annealing VECDA. The results achieve a given parallelism by choosing number of nodes, MPI processes per node and threads per MPI process. The number of threads increases as you move from left to right for given level of parallelism.

Note that the fluctuation effect is larger in the work reported here compared to our previous papers as we are looking here at many more simultaneous threads. Note that the effect does not just reflect the number of threads per process but also the total number of threads because the threads are synchronized not just within a process but between all processes as MPI calls will synchronize all the threads in the job. Thus it is interesting to examine this effect on the full 128 core Madrid cluster as this could even be a model for performance of future much larger core individual workstations.

We note that VECDA and PWDA differ somewhat in their performance characteristics. VECDA only uses modest size reductions (dominant use), broadcast and barrier MPI operations and so has particularly fast MPI synchronization. PWDA also has MPI\_SENDRECV (exchange of data between processes) which increases the MPI synchronization time. Thus VECDA shown in figures 18 and 19 tends always to have MPI at least as fast as CCR and in some cases very much faster. Figure 18 shows the parallel overhead for 44 different choices of nodes (from 1 to 8), MPI processes per node (from 1 to 16) and threads per node (from 1 to 16 divided between the MPI processes per node). The results are divided into groups corresponding to a given total parallelism. For each group, the number of threads increases as we move from left to right. For example in the 128 way parallel group, there are five entries with the leftmost

being 16 MPI processes per node on 8 nodes (a total of 128 MPI processes) and the rightmost 16 threads on each of 8 nodes (a total of 8 MPI processes). We find an incredibly efficient pure MPI version – an overhead of just 0.08 (efficiency 92%) for 128 way parallelism whereas the rightmost case of 16 threads has a 0.63 overhead (61% efficiency). All cases with 16 threads per node show a high overhead that slowly increases as the node count increases. For example the case of 16 threads on one node has an overhead of 0.51. Note that in this we use scaled speedup i.e. the problem size increases directly according to number of parallel units. This ensures that the inner execution scenarios are identical in all 44 cases reported in figure 18. We achieve scaled datasets by replicating a base point set as one can easily see that leads to same mathematical problem but with a work that increases properly as number of execution units increases.

Figure 19 looks again at the vector clustering VECDA comparing MPI versus two versions of threading. MPI is again very efficient – the 32 way parallel code with 16 MPI processes on each of two 16 core nodes has overheads (given by equation (1) and roughly 1 – efficiency) of 0.05 to 0.10. For the case of 16 threads on each of two nodes the overhead is 0.65 (short lived) to 1.25 (long lived) threads. The short lived threads are the natural implementation with threads spawned for parallel for loops. In the long lived case, the paradigm is similar to MPI with long running threads synchronizing with rendezvous semantics.

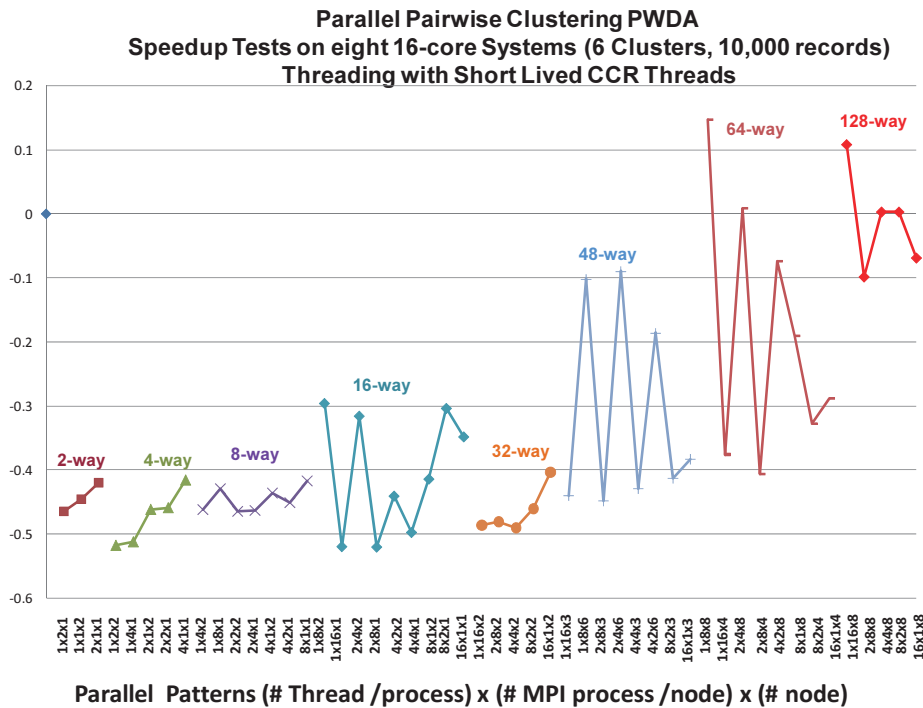


Figure 20. Parallel Overhead for PWDA runs on 128 core cluster (Ref. F in table 1) with patterns defined in figure 16.and in label in figure itself.

Figure 20 shows results of PWDA for a 10,000 element dataset on the 128 core cluster (ref. F in Table 1). The results show threading outperforming MPI for the highly

parallel results on right whereas on left (2- to 8-way parallelism) MPI outperforms threading. That is due to MPI being affected by the communication overhead of send-receive as discussed above for the results of figure 16. The results also show effects of the cache seen in the negative overheads (corresponding to a slow 1x1x1 case). The patterns are always labeled as (threads per process)x(MPI processes per node)x(nodes). Note figures 16 and 20 study the overhead for a fixed problem whereas figures 18 and 19 look at scaled speedup with problem size increasing proportional to number of parallel units. We see that the 10,000 element dataset can run well up even up to 128-way parallelism.

## 5. Conclusions

This paper has addressed several issues. It has studied the performance of a variety of different programming models on data intensive problems. It has presented novel clustering and MDS algorithms which are shown to parallelize well and could become supercomputer applications for large million point problems. It has compared MPI and threading on multicore systems showing both to be effective but with different overheads. We see these complemented by the data intensive programming models including Dryad and Hadoop as well as an in house version of MapReduce. These support an “owner stores and computes” programming paradigm that will be of increasing importance.

## Acknowledgements

We thank other members of the SALSA parallel computing group including Jong Choi and Yang Ruan for conversations. Our work on patient data would not have been possible with Gilbert Liu from IU Medical School and the gene sequence data was produced by Haixu Tang and Mina Rho from the IU Bioinformatics group who also answered our endless naïve questions. The IU MPI.NET group led by Andrew Lumsdaine was very helpful as we learnt how to use this. Our research was partially supported by Microsoft. We have obtained continued help on Dryad from Roger Barga and CCR from George Chrysanthakopoulos and Henrik Frystyk Nielsen. Scott Beason provided key support on visualization and set up of Clusters. The work on Particle Physics data analysis was a collaboration with Deep Web Technologies from New Mexico and Julian Bunn at Caltech.

## References

- [1] F. Damera, *SPMD model: past, present and future*, *Recent Advances in Parallel Virtual Machine and Message Passing Interface*, 8th European PVM/MPI Users' Group Meeting, Santorini/Thera, Greece, 2001.
- [2] MPI (Message Passing Interface), <http://www-unix.mcs.anl.gov/mpi/>
- [3] J. Dean and S. Ghemawat, *Mapreduce: Simplified data processing on large clusters*, *ACM Commun.*, vol. 51, Jan. 2008, pp. 107-113.
- [4] Geoffrey Fox and Marlon Pierce *Grids Challenged by a Web 2.0 and Multicore Sandwich* Special Issue of *Concurrency&Computation:Practice&Experience* on Seventh IEEE International Symposium on

- Cluster Computing and the Grid — CCGrid 2007, Keynote Talk Rio de Janeiro Brazil May 15 2007 <http://grids.ucs.indiana.edu/ptliupages/publications/CCGridDec07-Final.pdf>
- [5] Dennis Gannon and Geoffrey Fox, *Workflow in Grid Systems* Concurrency and Computation: Practice & Experience 18 (10), 1009-19 (Aug 2006), Editorial of special issue prepared from GGF10 Berlin.
- [6] Taverna Open Source Workflow code produced by OMII-OK <http://taverna.sourceforge.net/>
- [7] C. Chu, S. Kim, Y. Lin, Y. Yu, G. Bradski, A. Ng, and K. Olukotun. *Map-reduce for machine learning on multicore*. In B. Scholkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 281–288. MIT Press, Cambridge, MA, 2007.
- [8] Apache Hama Open Source project for MapReduce and matrix and machine learning algorithms <http://incubator.apache.org/hama/>
- [9] M. Isard, M. Budiu, Y. Yu, A. Birrell, and D. Fetterly, *Dryad: Distributed data-parallel programs from sequential building blocks*, European Conference on Computer Systems, March 2007.
- [10] Seung-Hee Bae *Parallel Multidimensional Scaling Performance on Multicore Systems* at workshop on Advances in High-Performance E-Science Middleware and Applications in Proceedings of eScience 2008 Indianapolis IN December 7-12 2008 [http://grids.ucs.indiana.edu/ptliupages/publications/eScience2008\\_bae3.pdf](http://grids.ucs.indiana.edu/ptliupages/publications/eScience2008_bae3.pdf)
- [11] Xiaohong Qiu, Geoffrey Fox, H. Yuan, Seung-Hee Bae, George Chrysanthakopoulos, Henrik Frystyk Nielsen *High Performance Multi-Paradigm Messaging Runtime Integrating Grids and Multicore Systems* September 23 2007 published in proceedings of eScience 2007 Conference Bangalore India December 10-13 2007 <http://grids.ucs.indiana.edu/ptliupages/publications/CCRSept23-07eScience07.pdf>
- [12] Xiaohong Qiu, Geoffrey C. Fox, Huapeng Yuan, Seung-Hee Bae, George Chrysanthakopoulos, Henrik Frystyk Nielsen *PARALLEL CLUSTERING AND DIMENSIONAL SCALING ON MULTICORE SYSTEMS* Invited talk at the 2008 High Performance Computing & Simulation Conference (HPCS 2008) In Conjunction With The 22nd EUROPEAN CONFERENCE ON MODELLING AND SIMULATION (ECMS 2008) Nicosia, Cyprus June 3 - 6, 2008; Springer Berlin / Heidelberg Lecture Notes in Computer Science Volume 5101/2008 "Computational Science: ICCS 2008" ISBN 978-3-540-69383-3 Pages 407-416 DOI: [http://dx.doi.org/10.1007/978-3-540-69384-0\\_46](http://dx.doi.org/10.1007/978-3-540-69384-0_46)
- [13] Xiaohong Qiu, Geoffrey C. Fox, Huapeng Yuan, Seung-Hee Bae, George Chrysanthakopoulos, Henrik Frystyk Nielsen *Performance of Multicore Systems on Parallel Data Clustering with Deterministic Annealing* ICCS 2008: "Advancing Science through Computation" Conference; ACC CYFRONET and Institute of Computer Science AGH University of Science and Technology Kraków, POLAND; June 23-25, 2008. Springer Lecture Notes in Computer Science Volume 5101, pages 407-416, 2008. DOI: [http://dx.doi.org/10.1007/978-3-540-69384-0\\_46](http://dx.doi.org/10.1007/978-3-540-69384-0_46)
- [14] Xiaohong Qiu, Geoffrey C. Fox, Huapeng Yuan, Seung-Hee Bae, George Chrysanthakopoulos, Henrik Frystyk Nielsen *Parallel Data Mining on Multicore Clusters* 7th International Conference on Grid and Cooperative Computing GCC2008 Shenzhen China October 24-26 2008 <http://grids.ucs.indiana.edu/ptliupages/publications/qiu-ParallelDataMiningMulticoreClusters.pdf>
- [15] Home Page for SALSA Project at Indiana University <http://www.infomall.org/salsa>.
- [16] Kenneth Rose, Eitan Gurewitz, and Geoffrey C. Fox *Statistical mechanics and phase transitions in clustering* Phys. Rev. Lett. 65, 945 - 948 (1990)
- [17] Rose, K. *Deterministic annealing for clustering, compression, classification, regression, and related optimization problems*, Proceedings of the IEEE Vol. 86, pages 2210-2239, Nov 1998
- [18] T Hofmann, JM Buhmann *Pairwise data clustering by deterministic annealing*, IEEE Transactions on Pattern Analysis and Machine Intelligence 19, pp1-13 1997
- [19] Apache Hadoop, <http://hadoop.apache.org/core/>
- [20] Microsoft Robotics Studio is a Windows-based environment that includes end-to-end Robotics Development Platform, lightweight service-oriented runtime, and a scalable and extensible platform. For details, see <http://msdn.microsoft.com/robotics/>
- [21] Georgio Chrysanthakopoulos and Satnam Singh *An Asynchronous Messaging Library for C#, Synchronization and Concurrency in Object-Oriented Languages (SCOOL)* at OOPSLA October 2005 Workshop, San Diego, CA.
- [22] S. Ghemawat, H. Gobioff, and S. Leung, *The Google file system*, Symposium on Operating Systems Principles 2003, pp 29–43, 2003.
- [23] Disco project, <http://discoproject.org/>
- [24] Erlang programming language, <http://www.erlang.org/>
- [25] S. Pallickara and G. Fox, "NaradaBrokering: A Distributed Middleware Framework and Architecture for Enabling Durable Peer-to-Peer Grids," *Middleware* 2003, pp. 41-61.
- [26] Y. Yu, M. Isard, D. Fetterly, M. Budiu, U. Erlingsson, P. Kumar and J. Currey, *DryadLINQ: A System for General-Purpose Distributed Data-Parallel Computing Using a High-Level Language* Symposium on Operating System Design and Implementation (OSDI), San Diego, CA, December 8-10, 2008.



- [27] J. B. MacQueen, *Some Methods for classification and Analysis of Multivariate Observations*, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, vol. 1, pp. 281-297.
- [28] Jaliya Ekanayake, Shrideep Pallickara, and Geoffrey Fox *Map-Reduce for Data Intensive Scientific Analyses* Proceedings of the IEEE International Conference on e-Science. Indianapolis. 2008. December 7-12 2008 <http://grids.ucs.indiana.edu/ptliupages/publications/ekanayake-MapReduce.pdf>
- [29] M. Svens'en. *GTM: The Generative Topographic Mapping*. PhD thesis, Neural Computing Research Group, Aston University, Birmingham, U.K., 1998.
- [30] T. Kohonen. *Self-Organizing Maps*. Springer-Verlag, Berlin, Germany, 2001.
- [31] J. B. Kruskal and M. Wish. *Multidimensional Scaling*. Sage Publications Inc., Beverly Hills, CA, U.S.A., 1978.
- [32] I. Borg and P. J. Groenen. *Modern Multidimensional Scaling: Theory and Applications*. Springer, New York, NY, U.S.A., 2005.
- [33] J. de Leeuw. *Applications of convex analysis to multidimensional scaling*. Recent Developments in Statistics, pages 133–145, 1977.
- [34] J. de Leeuw. *Convergence of the majorization method for multidimensional scaling*. Journal of Classification, 5(2):163–180, 1988.
- [35] Douglas Gregor and Andrew Lumsdaine. *Design and Implementation of a High-Performance MPI for C# and the Common Language Infrastructure*. Principles and Practice of Parallel Programming (PPOPP), pages 133-142, February 2008. ACM.
- [36] MPI.NET Home Page <http://www.osl.iu.edu/research/mpi.net>
- [37] Alkes L. Price, Eleazar Eskin and Pavel A. Pevzner. *Whole-genome analysis of Alu repeat elements reveals complex evolutionary history*. Genome Res. 2004 14: 2245-2252 DOI: <http://dx.doi.org/10.1101/gr.2693004>
- [38] Bell JF, Wilson JS, Liu GC. *Neighborhood greenness and 2-year changes in body mass index of children and youth*. Am J Prev Med. Dec 2008;35(6):547-553.
- [39] J. B. Kruskal. *Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis*. Psychometrika, 29(1):1–27, 1964.
- [40] Y. Takane, F. W. Young, and J. de Leeuw. *Nonmetric individual differences multidimensional scaling: an alternating least squares method with optimal scaling features*. Psychometrika, 42(1):7–67, 1977.
- [41] Hansjorg Klock, Joachim M. Buhmann *Multidimensional scaling by deterministic annealing* in Energy Minimization Methods in Computer Vision and Pattern Recognition, Eds Pelillo M. and Hancock E.R., Proc. Intl. Workshop EMMCVPR Venice Italy, Springer Lecture Notes in Computer Science 1223 ppg. 246-260 May 1997
- [42] Hansjorg Klock, Joachim M. Buhmann, *Data visualization by multidimensional scaling: a deterministic annealing approach*, Pattern Recognition 33 (2000) 651}669
- [43] Anthony J. Kearsley, Richard A. Tapia, Michael W. Trosset *The Solution of the Metric STRESS and SSTRESS Problems in Multidimensional Scaling Using Newton's Method*, technical report 1995.
- [44] Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). *Maximum-likelihood from incomplete data via the EM algorithm*. J. R. Statist. Soc. Ser. B (methodological), 39, 1–38.
- [45] Naonori Ueda and Ryohei Nakano *Deterministic annealing EM algorithm* Neural Networks Volume 11, Issue 2, 31 March 1998, Pages 271-282 [http://dx.doi.org/10.1016/S0893-6080\(97\)00133-0](http://dx.doi.org/10.1016/S0893-6080(97)00133-0)
- [46] Jong Youl Choi, Indiana University, Private Communication 2009.
- [47] Seung-Hee Bae, Indiana University, Private Communication 2009.