**Applicability of DryadLinq to Scientific Applications**

Content List

1. Introduction
2. Overview
   1. Dryad
   2. Google MapReduce
   3. MapReduce++
   4. Other related technologies (which we don’t have extensive performance results)
3. Performance and usability of applications using Dryad
   1. **EST** (Expressed Sequence Tag) sequence assembly program using DNA sequence assembly program software CAP3.
   2. **Pairwise Alu sequence alignment** using Smith Waterman dissimilarity computations followed by MPI applications for Clustering and MDS (Multi Dimensional Scaling)
   3. **HEP** Processsing large column of physics data using software Root and produce histogram results for data analysis.
   4. **Kmeans Clustering**
   5. Inhomogeneous data study
4. Analysis
   1. Dryad vs. Hadoop vs. MapReduce++ Programming Model
   2. Performance and Usability of Dryad (including setting up of Dryad)
5. Summary of key features of applications that suitable/not for Dryad
6. Reference
7. Introduction

Applying high level parallel runtimes to data/compute intensive applications is becoming increasingly common. The simplicity of the MapReduce programming model and the availability of open source MapReduce runtimes such as Hadoop, are attracting more users to the MapReduce programming model. Recently, Microsoft has released DryadLinq for academic use, allowing users to experience a new programming model and a runtime that is capable of performing large scale data/compute intensive analyses.

The goal of our study is to explore the applicability of DraydLinq to real scientific applications and compare its performance with other relevant parallel runtimes such as Hadoop. To achieve this goal we have developed a series of scientific applications using DryadLinq, namely, CAP3 DNA sequence assembly program , Pairwise ALU sequence alignment, High Energy Physics data analysis, and K-means Clustering . Each of these applications has unique requirements for parallel runtimes. For example, the HEP data analysis application requires ROOT data analysis framework to be available in all the compute nodes and in Pairwise ALU sequence alignment the framework must handle computing of distance matrix with hundreds of millions of points. We have implemented all these applications using DraydLinq and Hadoop, and used them to compare the performance of these two runtimes. CGL-MapReduce and MPI are used in applications where the contrast in performance needs to be highlighted.

In the sections that follow, we first present an overview of the different parallel runtimes we use in this analysis followed by a detailed discussion of the data analysis applications we developed. Here we discuss the mappings of parallel algorithms to the DraydLinq programming model and present performance comparisons with Hadoop implementations of the same applications. In section 4 we analyze DraydLinq’s programming model comparing it with other relevant technologies such as Hadoop and CGL-MapReduce. We also include a set of usability requirements for DryadLinq. We present our conclusions in section 5.

1. Overview

This section presents a brief introduction to a set of parallel runtimes we use our evaluations.

* 1. Microsoft DryadLinq

Dryad [4] is a distributed execution engine for coarse grain data parallel applications. Dryad considers computation tasks as directed acyclic graph (DAG)s where the vertices represent computation tasks and while the edges acting as communication channels over which the data flow from one vertex to another. In the HPC version of DryadLINQ the data is stored in (or partitioned to) Windows shared directories in local compute nodes and a meta-data file is use to produce a description of the data distribution and replication. Dryad schedules the execution of vertices depending on the data locality. (Note: The academic release of Dryad only exposes the DryadLINQ API for programmers. Therefore, all our implementations are written using DryadLINQ although it uses Dryad as the underlying runtime). Dryad also stores the output of vertices in local disks, and the other vertices which depend on these results, access them via the shared directories. This enables Dryad to re-execute failed vertices, a step which improves the fault tolerance in the programming model.

* 1. Apache Hadoop

Apache Hadoop has a similar architecture to Google’s MapReduce runtime , where it accesses data via HDFS, which maps all the local disks of the compute nodes to a single file system hierarchy, allowing the data to be dispersed across all the data/computing nodes. HDFS also replicates the data on multiple nodes so that failures of any nodes containing a portion of the data will not affect the computations which use that data. Hadoop schedules the MapReduce computation tasks depending on the data locality, improving the overall I/O bandwidth. The outputs of the *map* tasks are first stored in local disks until later, when the *reduce* tasks access them (pull) via HTTP connections. Although this approach simplifies the fault handling mechanism in Hadoop, it adds a significant communication overhead to the intermediate data transfers, especially for applications that produce small intermediate results frequently.

* 1. MapReduce++

MapReduce++ is a light-weight MapReduce runtime (earlier called CGL-MapReduce) that incorporates several improvements to the MapReduce programming model such as (i) faster intermediate data transfer via a pub/sub broker network; (ii) support for long running *map/reduce* tasks; and (iii) efficient support for iterative MapReduce computations. The use of streaming enables MapReduce++ to send the intermediate results directly from its producers to its consumers, and eliminates the overhead of the file based communication mechanisms adopted by both Hadoop and DryadLINQ. The support for long running *map/reduce* tasks enables configuring and re-using of *map/reduce* tasks in the case of iterative MapReduce computations, and eliminates the need for the re-configuring or the re-loading of static data in each iteration.

* 1. MPI

MPI , the de-facto standard for parallel programming, is a language-independent communications protocol that uses a message-passing paradigm to share the data and state among a set of cooperative processes running on a distributed memory system. MPI specification (F defines a set of routines to support various parallel programming models such as point-to-point communication, collective communication, derived data types, and parallel I/O operations. Most MPI runtimes are deployed in computation clusters where a set of compute nodes are connected via a high-speed network connection yielding very low communication latencies (typically in microseconds). MPI processes typically have a direct mapping to the available processors in a compute cluster or to the processor cores in the case of multi-core systems.. We use MPI as the baseline performance measure for the various algorithms that are used to evaluate the different parallel programming runtimes. Table 1 summarizes the different characteristics of Hadoop, Dryad, CGL-MapReduce, and MPI.

1. Comparison of features supported by different parallel programming runtimes.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Feature** | **Hadoop** | **DryadLINQ** | **MapReduce++** | **MPI** |
| Programming Model | MapReduce | DAG based execution flows | MapReduce with a  *Combine* phase | Variety of topologies constructed using the rich set of parallel constructs |
| Data Handling | HDFS | Shared directories/ Local disks | Shared file system / Local disks | Shared file systems |
| Intermediate Data Communication | HDFS/  Point-to-point via HTTP | Files/TCP pipes/ Shared memory FIFO | Content Distribution Network (NaradaBrokering (Pallickara and Fox 2003) ) | Low latency communication channels |
| Scheduling | Data locality/  Rack aware | Data locality/ Network  topology based run time graph optimizations | Data locality | Available processing capabilities |
| Failure Handling | Persistence via HDFS  Re-execution of map and reduce tasks | Re-execution of vertices | Currently not implemented  (Re-executing map tasks, redundant reduce tasks) | Program level  Check pointing  OpenMPI ,  FT MPI |
| Monitoring | Monitoring support of HDFS, Monitoring MapReduce computations | Monitoring support for execution graphs | Programming interface to monitor the progress of jobs | Minimal support for task level monitoring |
| Language Support | Implemented using Java. Other languages are supported via Hadoop Streaming | Programmable via C#  DryadLINQ provides LINQ programming API for Dryad | Implemented using Java  Other languages are supported via Java wrappers | C, C++, Fortran, Java, C# |

1. Performance and Usability of Applications using Dryad

In this section, we present the details of the DryadLINQ applications that we developed, the techniques we adopted in optimizing the applications, and their performance characteristics compared with Hadoop implementations. For our benchmarks, we used three clusters with almost identical hardware configurations with 256 CPU cores in each and a large cluster with 768 cores as shown in Table 2.

1. Different computation clusters used for this analysis.

| **Feature** | **Linux Cluster**  **(Ref A)** | **Windows Cluster (Ref B)** | **Windows Cluster (Ref C)** | **Windows Cluster (Ref D)** |
| --- | --- | --- | --- | --- |
| CPU | Intel(R) Xeon(R) CPU L5420 2.50GHz | Intel(R) Xeon(R) CPU L5420 2.50GHz | Intel(R) Xeon(R) CPU L5420 2.40GHz | Intel(R) Xeon(R) CPU L5420 2.50GHz |
| # CPU  # Cores | 2  8 | 2  8 | 4  6 | 2  8 |
| Memory | 32GB | 16 GB | 48 GB | 32 GB |
| # Disk | 1 | 2 | 1 | 1 |
| Network | Giga bit Ethernet | Giga bit Ethernet | Giga bit Ethernet | Giga bit Ethernet |
| Operating System | Red Hat Enterprise Linux Server release 5.3 -64 bit | Microsoft Window HPC Server 2008 (Service Pack 1) - 64 bit | Microsoft Window HPC Server 2008 (Service Pack 1) - 64 bit | Microsoft Window HPC Server 2008 (Service Pack 1) - 64 bit |
| # Cores | 256 | 256 | 768 | 256 |

* 1. EST (Expressed Sequence Tag) sequence assembly program using DNA sequence assembly program software CAP3.

CAP3 is a DNA sequence assembly program, developed by Huang and Madan [4], which performs several major assembly steps such as computation of overlaps, construction of contigs, construction of multiple sequence alignments and generation of consensus sequences, to a given set of gene sequences. The program reads a collection of gene sequences from an input file (FASTA file format) and writes its output to several output files and to the standard output as shown below. During an actual analysis, the CAP3 program is invoked repeatedly to process a large collection of input FASTA file.

*Input.fasta -> Cap3.exe -> Stdout + Other output files*

We developed a DryadLINQ application to perform the above data analysis in parallel. This application takes as input a *PartitionedTable* defining the complete list of FASTA files to process. For each file, the CAP3 executable is invoked by starting a process. The input collection of file locations is built as follows: (i) the input data files are distributed among the nodes of the cluster so that each node of the cluster stores roughly the same number of input data files; (ii) a “data partition” (A text file for this application) is created in each node containing the file paths of the original data files available in that node; (iii) a DryadLINQ “partitioned file” (a meta-data file understood by DryadLINQ) is created to point to the individual data partitions located in the nodes of the cluster.

Following the above steps, a DryadLINQ program can be developed to read the data file paths from the provided partitioned-file, and execute the CAP3 program using the following two lines of code.

IQueryable<Line Record> filenames = PartitionedTable.Get<LineRecord>(uri);

IQueryable<int> exitCodes= filenames.Select(s => ExecuteCAP3(s.line));

Although we use this program specifically for the CAP3 application, the same pattern can be used to execute other programs, scripts, and analysis functions written using the frameworks such as R and Matlab, on a collection of data files. (Note: In this application, we rely on DryadLINQ to process the input data files on the same compute nodes where they are located. If the nodes containing the data are free during the execution of the program, the DryadLINQ runtime will schedule the parallel tasks to the appropriate nodes to ensure co-location of process and data; otherwise, the data will be accessed via the shared directories.)

* + 1. Evaluations and Findings

CAP3 is a compute intensive application that operates on a comparably smaller data sets (thousands of FASTA files but the size in disk is small such as few gigabytes). When we first deployed the application on the cluster, we noticed a sub-optimal CPU utilization, which seemed highly unlikely for this pleasingly parallel a compute intensive application. A trace of job scheduling in the HPC cluster revealed that the scheduling of individual CAP3 executables in a given node was not always utilizing all CPU cores. We traced this behavior to the use of an early version of the PLINQ [12] library (June 2008 Community Technology Preview), which DryadLINQ uses to achieve core level parallelism on a single machine. During this analysis we identified three scheduling issues (including the above) related to DryadLINQ and the software it uses. They are:

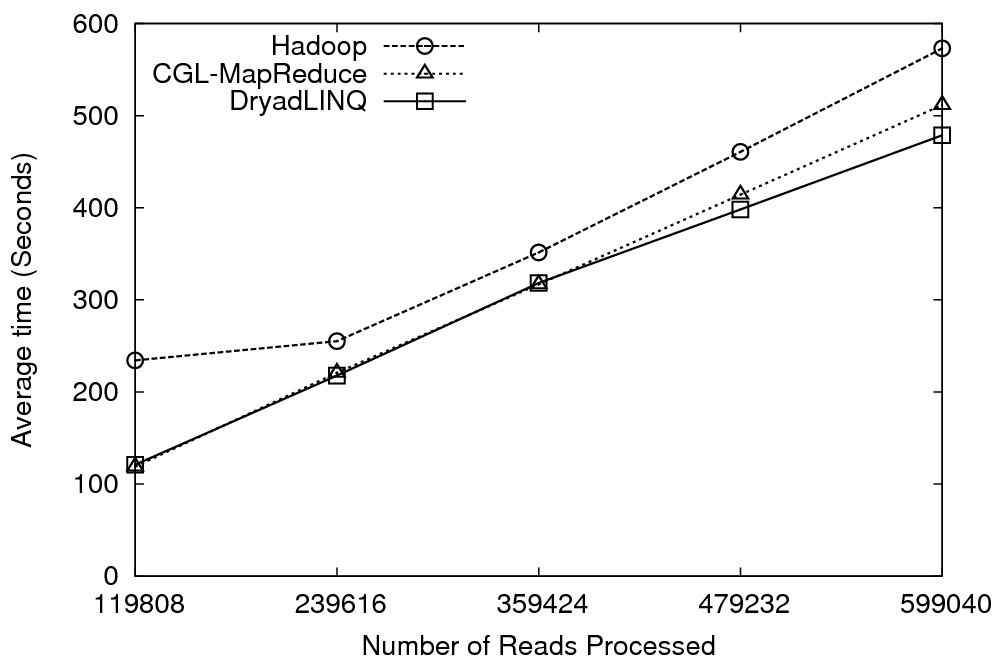
1. DryadLINQ schedule jobs to nodes rather than cores – ideal cores when the data is inhomogeneous.
2. PLINQ does not utilize CPU cores well.
3. Performance of threads is extremely low for memory intensive operations compared to processes.

The cause for the lower utilization of CAP3 application is due to the PLINQ’s (June 2008 Community Technology Preview) unoptimized scheduling mechanism. For example, we have observed a scheduling of 8 ->4 ->4 parallel tasks in an 8 CPU core node for 16 parallel tasks, which should ideally be scheduled as two batches of 8 parallel tasks (8->8).

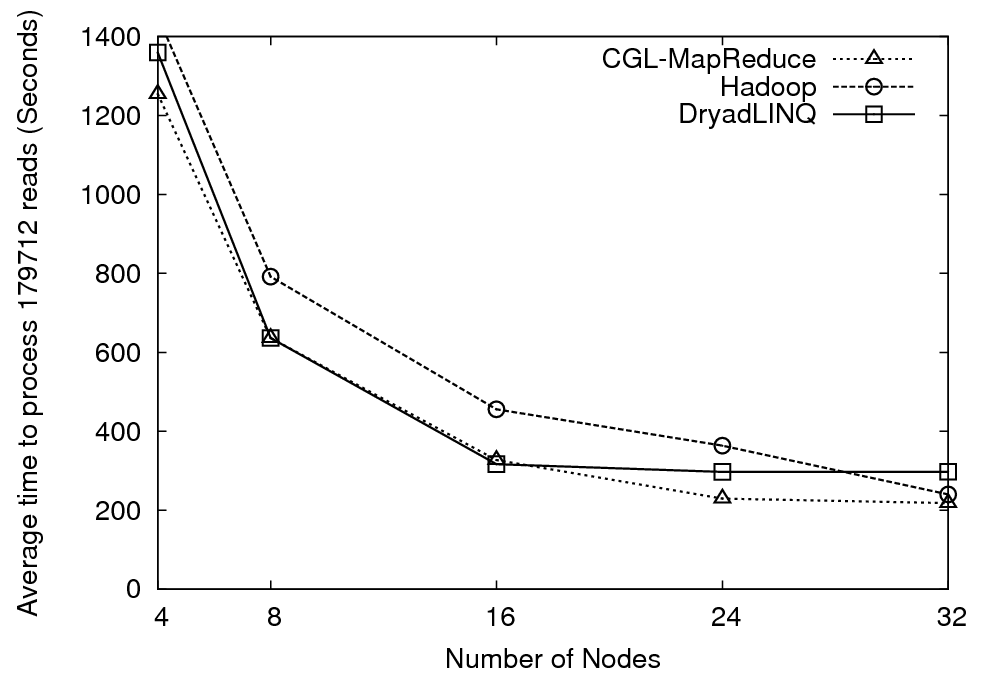
**We verified that this issue has been fixed in the current version of PLINQ and future releases of DryadLINQ will benefit from these improvements.**

However, the users of the academic release of DryadLINQ would experience the above issue and most intuitive programs written using DryadLINQ may not utilize CPU cores well. While using the preview version of PLINQ (which is publically available), we were able to reach full CPU utilization using the Academic release of DryadLINQ by changing the way we partition the data. Instead of partitioning input data to a single data-partition per node, we created data-partitions containing at most 8 (=number of CPU cores) line records (actual input file names). This way, we used DryadLINQ’s scheduler to schedule series of vertices corresponding to different data-partitions in nodes while PLINQ always schedules 8 tasks at once, which gave us 100% CPU utilization.

We evaluated DryadLINQ and Hadoop for the CAP3 application using the above optimized DryadLINQ program. Figure 1 and 2 show comparisons of performance and the scalability of the DryadLINQ application, with the Hadoop and CGL-MapReduce versions of the CAP3 application.



1. Performance of different implementations of CAP3 application.



1. Scalability of different implementations of CAP3.

The performance and the scalability graphs shows that all three runtimes work almost equally well for the CAP3 program, and we would expect them to behave in the same way for similar applications with simple parallel topologies. Except for the manual data partitioning requirement, implementing this type of applications using DryadLinq is extremely simple and straightforward.

* + 1. Inhomogeneity of data partitions and scheduling partitions to nodes

DryadLINQ schedules vertices of the DAG (corresponding to data partitions) to compute nodes rather than individual CPU cores. This may also produce suboptimal CPU utilizations of Dryad programs depending on the data partition strategy. As in MapReduce programming model, Dryad also assumes that the vertices corresponding to a given phase of the computation partitions data so that the data is distributed evenly across the computation nodes. Although this is possible in some computations such as sorting and histogramming where the data can be divisible arbitrary, it is not always possible when there are inhomogeneous data products at the lowest level of the data items such as gene sequences, binary data files etc.. For example, CAP3 process sequence data as a collection of FASTA files and the number of sequences contaminating in each of these files may differ significantly causing imbalanced workloads.

Since DryadLINQ schedules vertices to nodes, it is possible that a vertex which processes few large FASTA files using few CPU cores of a compute node will keep all the other CPU cores of that machine idle. In Hadoop, the map/reduce tasks are scheduled to individual CPU cores (customized by the user) and hence it is able to utilize all the CPU cores to execute map/reduce tasks in a given time.

E:\academic\phd\Publications\MTIGS09\gnuplot\CAP3_active_workers_bw.eps

1. Number of active tasks/CPU cores along the running times of two runs of CAP3.

The collection of input files we used for the benchmarks contained different number of gene sequences in each, and hence it did not represent a uniform workload across the concurrent vertices of the DryadLINQ application, because the time the CAP3 takes to process an input file varies depending on the number of sequences available in the input file. The above characteristics of the data produces lower efficiencies at higher number of CPU cores as more CPU cores become idle towards the end of the computation waiting for vertices that takes longer time to complete.

To verify the above observation we measured the utilization of vertices during two runs of the CAP3 program. In our first run we used 768 input files so that Dryad schedules 768 vertices on 768 CPU cores, while in the second Dryad schedules 1536 vertices on 768 CPU cores. The result of this benchmark is shown in figure 3. The first graph in figure 3 corresponding to 768 files indicates that although DryadLINQ starts all the 768 vertices at the same time they finish at different times with long running tasks taking roughly 40% of the overall time. The second graph (1536 files) shows that the above effect has caused lower utilization of vertices when Dryad schedules 1536 vertices to 768 CPU cores. This is due to the way Dryad schedule tasks to compute nodes and we have discussed this behavior extensively in [25].

1. Cap3 Dryad implementation performance comparison of data sets with large standard deviation & 0 standard deviation.

In this test we chose two data sets with similar mean file size values (444.5 kB), but with different standard deviations. The first data set of 1152 files was created by replicating a single file, making the standard deviation 0. The second data set contained 1152 files of different sizes with a standard deviation of 185kB. The results are shown in figure 4, where it’s clearly evident that the data set with a large standard deviation takes more time to execute. This further confirms the observation we made above. A more thorough study is currently under way to understand the effects of data inhomogeneity in cap3 implementations.

* 1. Pairwise Alu sequence alignment using Smith Waterman dissimilarity computations followed by MPI applications for Clustering and MDS (Multi Dimensional Scaling)
     1. ALU Clustering

The ALU clustering problem [26] is one of the most challenging problems for sequencing clustering because ALUs represent the largest repeat families in human genome. There are about 1 million copies of ALU sequences in human genome, in which most insertions can be found in other primates and only a small fraction (~ 7000) are human-specific. This indicates that the classification of ALU repeats can be deduced solely from the 1 million human ALU elements. Notable, ALU clustering can be viewed as a classical case study for the capacity of computational infrastructures because it is not only of great intrinsic biological interests, but also a problem of a scale that will remain as the upper limit of many other clustering problem in bioinformatics for the next few years, e.g. the automated protein family classification for a few millions of proteins predicted from large metagenomics projects.

* + 1. Smith Waterman Dissimilarities

We identified samples of the human and Chimpanzee ALU gene sequences using Repeatmasker [27] with Repbase Update [28]. We have been gradually increasing the size of our projects with the current largest samples having 35339 and 50000 sequences and these require a modest cluster such as Tempest (768 cores) for processing in a reasonable time (a few hours as shown in section 5). Note from the discussion in section 4.4.1, we are aiming at supporting problems with a million sequences -- quite practical today on TeraGrid and equivalent facilities given basic analysis steps scale like O(N2).

We used open source version NAligner [29] of the Smith Waterman – Gotoh algorithm SW-G [30,31] modified to ensure low start up effects by each thread/processing large numbers (above a few hundred) at a time. Memory bandwidth needed was reduced by storing data items in as few bytes as possible.

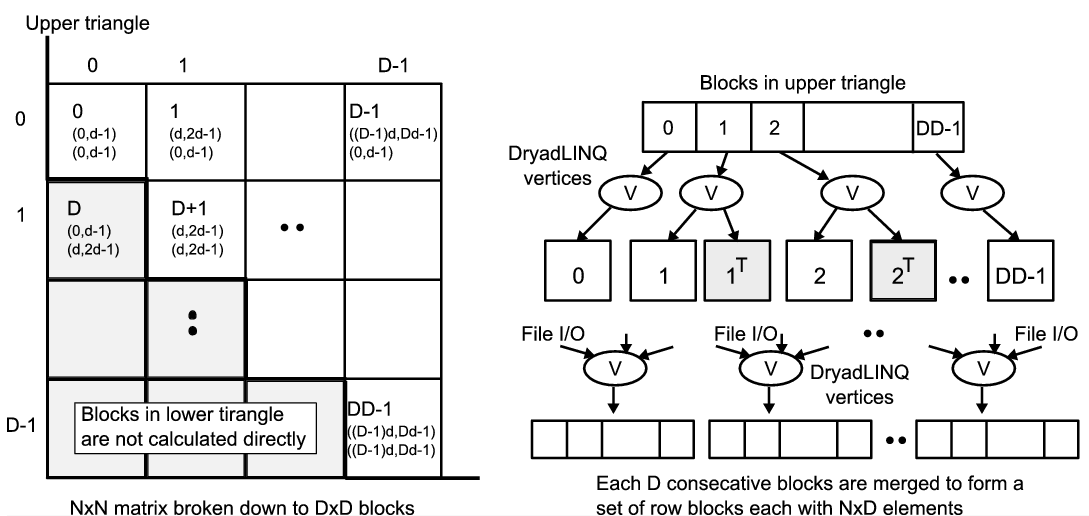
* + 1. The O(N2) Factor of 2 and structure of processing algorithm

The ALU sequencing problem shows a well known factor of 2 issue present in many O(N2) parallel algorithms such as those in direct simulations of astrophysical stems. We initially calculate in parallel the Distance D(i,j) between points (sequences) i and j. This is done in parallel over all processor nodes selecting criteria i < j (or j > i for upper triangular case) to avoid calculating both D(i,j) and the identical D(j,i). This can require substantial file transfer as it is unlikely that nodes requiring D(i,j) in a later step will find that it was calculated on nodes where it is needed.

For example the MDS and PW (PairWise) Clustering algorithms described in Fox, Bae et al. (2008), require a parallel decomposition where each of N processes (MPI processes, threads) has 1/N of sequences and for this subset {i} of sequences stores in memory D({i},j) for all sequences j and the subset {i} of sequences for which this node is responsible. This implies that we need D (i,j) and D (j,i) (which are equal) stored in different processors/disks. This is a well known collective operation in MPI called either gather or scatter.

* + 1. Dryad Implementation

We developed a DryadLINQ application to perform the calculation of pairwise SW-G distances for a given set of genes by adopting a coarse grain task decomposition approach which requires minimum inter-process communicational requirements to ameliorate the higher communication and synchronization costs of the parallel runtime. To clarify our algorithm, let’s consider an example where N gene sequences produces a pairwise distance matrix of size NxN. We decompose the computation task by considering the resultant matrix and groups the overall computation into a block matrix of size DxD where D is a multiple (>2) of the available computation nodes. Due to the symmetry of the distances D(i,j) and D(j,i) we only calculate the distances in the blocks of the upper triangle of the block matrix as shown in figure 5 (left). The blocks in the upper triangle are partitioned (assigned) to the available compute nodes and an “Apply” operation is used to execute a function to calculate (N/D)x(N/D) distances in each block. After computing the distances in each block, the function calculates the transpose matrix of the result matrix which corresponds to a block in the lower triangle, and writes both these matrices into two output files in the local file system. The names of these files and their block numbers are communicated back to the main program. The main program sort the files based on their block number s and perform another “Apply” operation to combine the files corresponding to a row of blocks in a single large row block as shown in the figure 5 (right).



1. Task decomposition (left) and the DryadLINQ vertex hierarchy (right) of the DryadLINQ implementation of SW-G pairwise distance calculation application.
   * 1. MPI Implementation

The MPI version of SW-G calculates pairwise distances using a set of either single or multi-threaded processes. For N gene sequences, we need to compute half of the values (in the lower triangular matrix), which is a total of M = N x (N-1) /2 distances. At a high level, computation tasks are evenly divided among P processes and execute in parallel. Namely, computation workload per process is M/P. At a low level, each computation task can be further divided into subgroups and run in T concurrent threads. Our implementation is designed for flexible use of shared memory multicore system and distributed memory clusters (tight to medium tight coupled communication technologies such threading and MPI). We provide options for any combinations of thread vs. process vs. node as shown in figure 6. The real computation workload per parallel unit is decided by M / (T x P x # nodes).

### Performance of Smith Waterman Gotoh SW-G Algorithm

We performed the Dryad and MPI implementations of ALU SW-G distance calculations on two large data sets and obtained the following results.

Table 4 Comparison of Dryad and MPI technologies on ALU sequencing application with SW-G algorithm

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Technology** | | **Total Time (seconds)** | **Time per Pair (milliseconds)** | **Partition Data (seconds)** | **Calculate and Output Distance(seconds)** | **Merge files (seconds)** |
| Dryad | 50,000 sequences | 17200.413 | 0.0069 | 2.118 | 17104.979 | 93.316 |
| 35,339 sequences | 8510.475 | 0.0068 | 2.716 | 8429.429 | 78.33 |
| MPI | 50,000 sequences | 16588.741 | 0.0066 | N/A | 13997.681 | 2591.06 |
| 35,339 sequences | 8138.314 | 0.0065 | N/A | 6909.214 | 1229.10 |

There is a short partitioning phase for Dryad and then both approaches calculate the distances and write out these to intermediate files as discussed in section 4. We note that merge time is currently much longer for MPI than Dryad while the initial steps are significantly faster for MPI. However the total times in table 4 indicates that both MPI and Dryad implementations perform well for this application with MPI a few percent faster with current implementations. As expected, the times scale proportionally to the square of the number of distances. On 744 cores the average time of 0.0067 milliseconds per pair that corresponds to roughly 5 milliseconds per pair calculated per core used. The coarse grained Dryad application performs competitively with the tightly synchronized MPI application. It proves once more the applicability of the Cloud technologies for the composable applications.

* + 1. Threaded Implementation

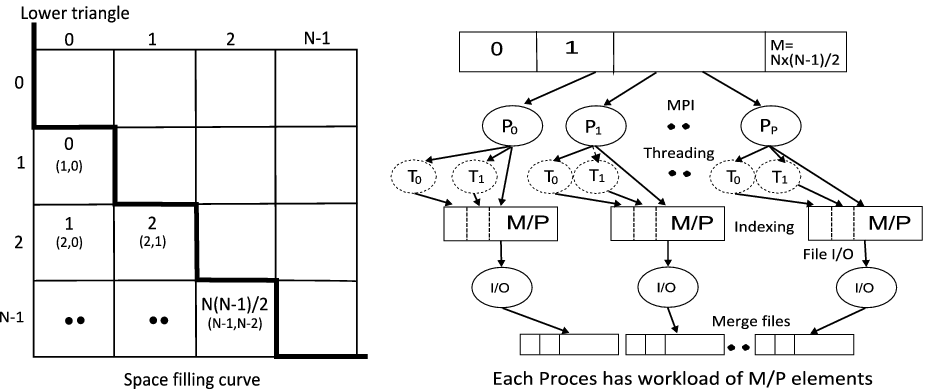
Above we looked at using MPI with one process per core and we compared this with a threaded implementation with each process having several threads. Labeling configuration as t x m x n for t threads per process, m MPI processes per node and n nodes, we compare choices of t, m and n in figure 6.

alu_mpi_bw_nokey_1.eps

1. Performance of ALU Gene Alignments for different parallel patterns

The striking result for this step is that MPI easily outperforms the equivalent threaded version of this embarrassingly parallel step. In figure 6, all the peaks in the overhead correspond to patterns with large values of t. Note that the MPI intra-node 1x24x32 pattern completes the full 624 billion alignments in 2.33 hours – 4.9 times faster than threaded implementation 24x1x32. This 768 core MPI run has a parallel overhead of 1.43 corresponding to a speed up of 316.

The SW-G alignment performance is probably dominated by memory bandwidth issues, and we are pursuing several points that could affect this though it is not at our highest priority as SW-G is not the dominant step. We have tried to identify the reason for the comparative slowness of threading. Using Windows monitoring tools, we found that the threaded version has about a factor of 100 more context switches than in the one thread per process MPI version. This could lead to a slow down of the threaded approach and correspond to Windows handing of paging of threads with large memory footprints.



1. Task decomposition (left) and the MPI (right) implementation of SW-G pairwise distance calculation application.

As illustrated in figure 7, the data decomposition strategy runs a "space filling curve through lower triangular matrix" to produce equal numbers of pairs for each parallel unit such as process or thread. It is necessary to map indexes in each pairs group back to corresponding matrix coordinates (i, j) for constructing full matrix later on. We implemented a special function "PairEnumertator" as the convertor. We tried to limit runtime memory usage for performance optimization. This is done by writing a triple of i, j, and distance value of pairwise alignment to a stream writer and the system flashes accumulated results to a local file periodically. As the final stage, individual files are merged to form a full distance matrix.

* + 1. Apache Hadoop Implementation

We developed an Apache Hadoop version of the pairwise distance calculation program based on the JAligner[26] program, the java implementation of the NAligner. Similar to the other implementations, the computation is partitioned in to blocks based on the resultant matrix. Each of the blocks would get computed as a map task. The block size (D) can be specified via an argument to the program. The block size needs to specified in such a way that there will be much more map tasks than the map task capacity of the system, so that the Apache Hadoop scheduling will happen as a pipeline of map tasks resulting in global load balancing of the application. The input data is distributed to the worker nodes through the Hadoop distributed cache, which makes them available in the local disk of each compute node.

A load balanced task partitioning strategy according to the following rules is used to identify the blocks that need to be computed (green) through map tasks as shown in the figure 8. In addition all the blocks in the diagonal (blue) are computed. Even though the task partitioning mechanisms are different, both Dryad-SWG and Hadoop-SWG ends up with essentially identical computation blocks, if the same block size is given to both the programs.

*If* β *>=* α*, we only calculate D(*α*,*β*) if* α*+*β *is even,  
 If* β *<* α*, we only calculate D(*α*,*β*) if* α*+*β *is odd.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **1**  (1-100) | **2**  (101-200) | **3**  (201-300) | **4**  (301-400) |
| **1**  (1-100) | **M1** | **M2** | *from M6* | **M3** | **Reduce 1 hdfs://.../rowblock\_1.out** |
| **2**  (101-200) | *from M2* | **M4** | **M5** | *from M9* | **Reduce 2**  **hdfs://.../rowblock\_2.out** |
| **3**  (201-300) | **M6** | *from M5* | **M7** | **M8** | **Reduce 3**  **hdfs://.../rowblock\_3.out** |
| **4**  (301-400) | *from M3* | **M9** | *from M8* | **M10** | **Reduce 4**  **hdfs://.../rowblock\_4.out** |

1. Task (Map) decomposition and the reduce task data collection example.

Map tasks use custom Hadoop writable objects as the map task output values to store the calculated pairwise distance matrices for the respective blocks. In addition, non-diagonal map tasks output the inverse distances matrix as a separate output value. Hadoop uses local files and http transfers underneath to transfer the map task output key value pairs to the reduce tasks.

The outputs of the map tasks are collected by the reduce tasks. Since the reduce tasks start collecting the outputs as soon as the first map task finishes and continue to do so while other map tasks are executing, the data transfers from the map tasks to reduce tasks do not present a significant performance overhead to the program. The program currently creates a single reduce task per each row block resulting in total of (no. of sequences/block size) Reduce tasks. Each reduce task to accumulate the output distances for a row block and writes the collected output to a single file in Hadoop Distributed File System (HDFS). This results in N number of output files corresponding to each row block, similar to the output we produce in the Dryad version.

### Performance comparison of Dryad and Hadoop implementations

We compared the Dryad and Hadoop implementations on the same data sets we used for the Dryad and MPI comparisons, but on a different cluster. These tests were run on cluster ref A for Hadoop-swg and on ref D for Dryad-swg, which are two identical Linux and Windows clusters. The Dryad-adjusted results row represents the performance timings adjusted for the performance difference of the base programs, NAligner and the JAligner. In here we do not present separate times for the merge step as the Hadoop implementation performs the merging with reduce tasks even when the map tasks are running.

Table 5 Comparison of Dryad and Hadoop technologies on ALU sequencing application with SW-G algorithm

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Technology | No. of Sequences | Total Time  (seconds) | Time per  Pair (ms) | No of actual  Alignments | Sequential  Time (seconds) | Speedup | Speedup  per core |
| Dryad | 50,000 Sequences | 30881.74 | 0.0124 | 1259765625 | 4884111.33 | 158.16 | 61.78% |
| 35,339 | 14654.41 | 0.0117 | 634179061 | 2458712.22 | 167.78 | 65.54% |
| Dryad –adjusted | 50,000 | 24202.4 | 0.0097 | 1259765625 | 3827736.66 | 158.16 | 61.78% |
| 35,339 | 11484.84 | 0.0092 | 634179061 | 1926922.27 | 167.78 | 65.54% |
| Hadoop | 50,000 | 17798.59 | 0.0071 | 1262500000 | 4260937.50 | 239.40 | 93.51% |
| 35,339 | 8974.638 | 0.0072 | 629716021 | 2125291.57 | 236.81 | 92.50% |

We can notice that the Hadoop implementation shows more speedup per core than the Dryad implementation. In an alternate ongoing testing we are noticing that the block size plays a larger role with regards to the dryad implementation performance, where relatively smaller block sizes are performing better. This led us to speculate that the lower speedup of Dryad implementation is related to the memory usage. We are currently pursuing this issue more deeply to understand the reasons for this behavior.

* + 1. Inhomogeneous data study

The time complexity to align and obtain distances for two genome sequences with lengths ‘m’ and ‘n’ using SW-G algorithm is proportional to the product of the lengths of two sequences, O(mn). This makes the sequence length distribution of a block to determine the execution time for that particular execution block. Frameworks like Dryad and Hadoop work optimally when the work is equally partitioned among the tasks, striving for equal length sequences in the case of pairwise distance calculations. Depending on the scheduling strategy of the framework, blocks with different execution times can have a adverse effect on the performance of the applications, unless proper load balancing measures have been taken in the task partitioning steps. For an example, in Dryad vertices are scheduled at the node level, making it possible for a node to have blocks with varying execution times. In this case if a single block inside a vertex takes a larger amount of time than other blocks to execute, then the whole node have to wait till the large task completes, which utilizes only a fraction of the node resources.

Sequence sets that we encounter in the real data sets are inhomogeneous in length. In this section we study the effect of inhomogeneous gene sequence lengths for our pairwise distance calculation applications. The data sets used were randomly generated with a given mean sequence length (400) with varying standard deviations following a normal distribution of the sequence lengths. Each data set contained a set of 10000 sequences, 100 million pairwise distance calculations to perform. Sequences of varying lengths are randomly distributed across the data set.

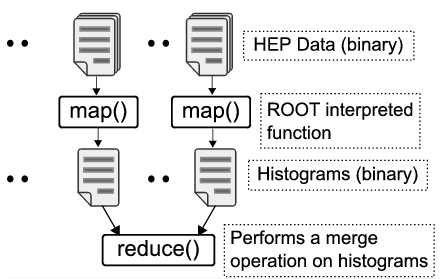
1. Performance of SW-G pairwise distance calculation application for inhomogeneous data

The Dryad-adjusted results depict the raw Dryad results adjusted for the performance difference of the NAligner and JAligner base programs. As we notice from the figure 9, both Dryad implementation as well as the Hadoop implementation performed satisfactorily, without showing significant performance degradations. In fact Hadoop implementation showed minor improvements in the execution times. The acceptable performance can be attributed to the fact that the sequences with varying lengths are randomly distributed across the data set, giving a natural load balancing to the sequence blocks. We are currently working studying the effect of different inhomogeneous data distributions, specially the skewed distributions like a set of sequences sorted on the lengths, on the performance of different implementations to develop data partition strategies to mitigate the potential pitfalls.

The Hadoop implementations’ slight performance improvement can be attributed to the global pipeline scheduling of map tasks that Hadoop performs. In Hadoop administrator can specify the map task capacity of a particular worker node and then Hadoop global scheduler schedules the map tasks directly on to those placeholders in a much finer granularity than in Dryad as an when individual tasks finish. This allows the Hadoop implementation to perform natural global level load balancing. In this case it might even be advantageous to have varying task execution times to iron out the effect of any trailing map tasks towards the end.

* 1. HEP Processing large column of physics data using software Root and produce histogram results for data analysis.

HEP data analysis application has a typical MapReduce application in which the *map* phase is used process a large collection of input files containing events (features) generated by HEP experiments. The output of the *map* phase is a collection of partial histograms containing identified features. During the reduction phase these partial histograms are merged to produce a single histogram representing the overall data analysis. Figure 10 shows the data flow of the HEP data analysis application.



1. Program/data flow of the HEP data analysis application

Although the structure of this application is simple and fits perfectly well with the MapReduce programming model, it has a set of highly specific requirements such as:

1. All data processing functions are written using an interpreted language supported by ROOT data analysis framework
2. All the data products are in binary format and passed as files to the processing scripts
3. Large input data sets (Large Hadron Collider will produce 15 petabytes of data per year).

We manually partitioned the input data to the compute nodes of the cluster and generated data-partitions containing only the file names available in a given node. The first step of the analysis requires applying a function coded in ROOT to all the input files. The analysis script we used can process multiple input files at once, therefore we used a homomorphic Apply (shown below) operation in DryadLINQ to perform the first stage (corresponding to the *map()* stage in MapReduce) of the analysis.

[Homomorphic]

ApplyROOT(string fileName){..}

IQueryable<HistoFile> histograms = dataFileNames.Apply(s => ApplyROOT (s));

Unlike the Select operation that processes records one by one, the Apply operation allows a function to be applied to an entire data set, and produce multiple output values. Therefore, in each vertex the program can access a data partition available in that node (provided that the node is available for executing this application – please refer to the “Note” under CAP3 section). Inside the ApplyROOT() method, the program iterates over the data set and groups the input data files, and execute the ROOT script passing these files names along with other necessary parameters. The output of this operation is a binary file containing a histogram of identified features of the input data. The ApplyROOT() method saves the output histograms in a predefined shared directory and produces its location as the return value.

In the next step of the program, we perform a combining operation of these partial histograms. Again, we use a homomorphic Apply operation to combine partial histograms. Inside the function that is applied to the collection of histograms, we use another ROOT script to combine collections of histograms in a given data partition. (Before this step, the main program generates the data-partitions containing the histogram file names). The output partial histograms produced by the previous step will be combined by the main program to produce the final histogram of identified features.

* + 1. Evaluations and Findings

The first task we had to tackle in the DryadLinq implementation of this application is the distribution of data across the computation cluster. We used a data set of one terabytes (1TB) and hence storing and distributing this data set poses challenges. Typically these large data sets are stored in shared file systems and then get distributed to the computation nodes before the analysis. In this application the input data is organized in a large number of binary files each of which roughly occupy 33MB of disk space. Distributing a collection of data files across a computation cluster is a typical requirement in many scientific applications and we have already experienced this in CAP3 data analysis as well.

Current release of DryadLinq does not provide any tools to do such data distribution. However, it provides two partitioning constructs which can be use to develop an application to perform this data distribution. One possible approach is to develop a DraydLinq application to copy input files from its shared repository to individual computation units. This may saturate the shared repository infrastructure as all the compute nodes try to copy data from this shared location. We developed a standalone application to perform the above distribution as it can be used for many similar situations.

Hadoop provides and optimized solution to distributing data across computation nodes of a cluster via HDFS and a client tool. The above data distribution reduces to the following simple command in the Hadoop environment.

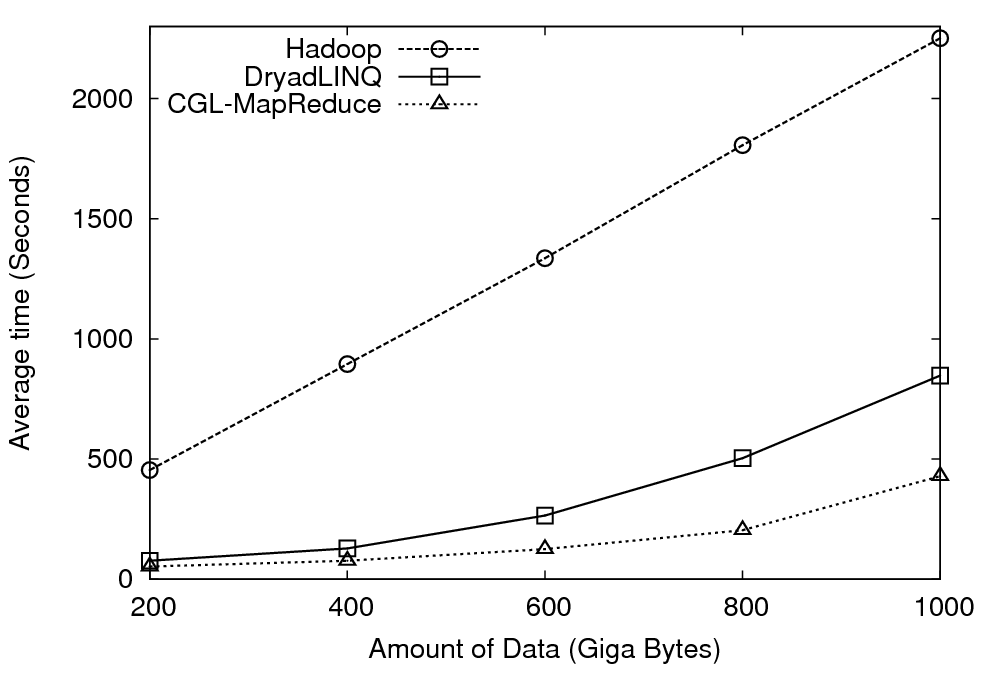
bin/Hadoop –dfs put shared\_repository\_path destination\_in\_hdfs

We think that a similar tool for DryadLinq would help users to partition data (available in files) more easily than developing custom solution for each application.

The second challenge we faced in implementing the above application is the use of ROOT data analysis framework to process data. This is also a common requirement in may scientific analysis as many data analysis functions are written using specific analysis software such as ROOT, R, Matlab etc. To use these specific software at DryadLinq vertices they need to be installed in each and every compute node of the cluster. Some of these applications only require copying a collection of libraries to the compute nodes while some requires complete installations. Clusrun is a possible solution to handle both types of installations, however providing another simple tool to perform the first type of installations would benefit the users. (Note: we could ship few shared libraries or other necessary resources using DryadLinq.Resources.Add(resource\_name) method. However, this does not allow user to add a folder of libraries or a collection of folders. The ROOT installation requires copying few folders to every compute node)

After tackling the above two problems we were able to develop a DryadLinq application for the HEP data analysis. As in CAP3 program, we noticed sub-optimal utilization of CPU cores by the HEP application due to the above mention problem in the early version of PLINQ (June 2008 CTP). With heterogeneous processing times of different input files, we were able to correct this partially by carefully selecting the number of data partitions and the amount of records accessed at once by the ApplyROOT() function.

We measure the performance of this application with different input sizes up to 1TB of data and compare the results with Hadoop and CGL-MapReduce implementations that we have developed previously. The results of this analysis are shown in Figure 11.



1. Performance of different implementations of HEP data analysis applications.

The results in Figure 11 highlight that Hadoop implementation has a considerable overhead compared to DraydLINQ and CGL-MapReduce implementations. This is mainly due to differences in the storage mechanisms used in these frameworks. DryadLINQ and CGL-MapReduce access the input from local disks where the data is partitioned and distributed before the computation. Currently, HDFS can only be accessed using Java or C++ clients, and the ROOT – data analysis framework is not capable of accessing the input from HDFS. Therefore, we placed the input data in IU Data Capacitor – a high performance parallel file system based on Lustre file system, and allowed each map task in Hadoop to directly access the input from this file system. This dynamic data movement in the Hadoop implementation incurred considerable overhead to the computation. In contrast, the ability of reading input from the local disks gives significant performance improvements to both Dryad and CGL-MapReduce implementations.

Additionally, in the DryadLINQ implementation, we stored the intermediate partial histograms in a shared directory and combined them during the second phase as a separate analysis. In Hadoop and CGL-MapReduce implementations, the partial histograms are directly transferred to the *reducers* where they are saved in local file systems and combined. These differences can explain the performance difference between the CGL-MapReduce version and the DryadLINQ version of the program. We are planning to develop a better version of this application for DryadLINQ in the future.

* 1. K-means Clustering

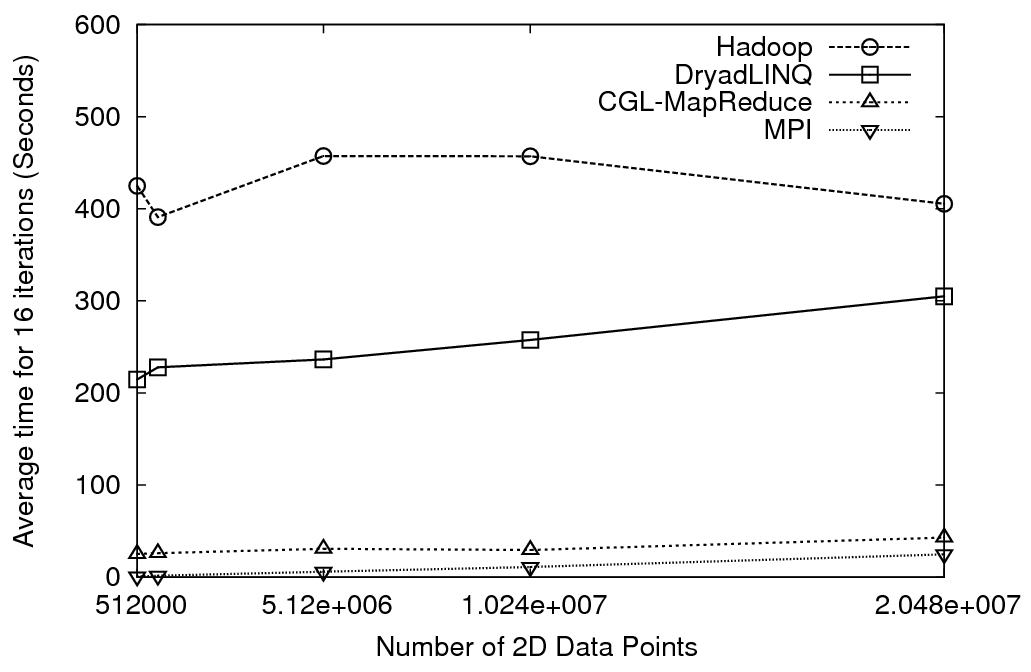
We implemented a K-means Clustering application using DryadLINQ to evaluate its performance under iterative computations. Algorithms such as clustering, matrix multiplication, Multi Dimensional Scaling are some examples that performs iterative computations. We used K-means clustering to cluster a collection of 2D data points (vectors) to a given number of cluster centers. The MapReduce algorithm we used is shown below. (Assume that the input is already partitioned and available in the compute nodes). In this algorithm, *Vi* refers to the *ith* vector, *Cn,j* refers to the *jth* cluster center in *nth* iteration, *Dij* refers to the Euclidian distance between *ith* vector and *jth* cluster center, and *K* is the number of cluster centers.

|  |
| --- |
| **K-means Clustering Algorithm for MapReduce** |
| **Do**  Broadcast *Cn*  *[Perform in parallel] –the map() operation*  **for each** *V*i  **for each** *Cn,j*  *Dij <= Euclidian (Vi,Cn,j)*  Assign point *Vi* to *Cn,j* with minimum *Dij*  **for each** *Cn,j*  *Cn,j <=Cn,j/K*  *[Perform Sequentially] –the reduce() operation*  Collect all *Cn*  Calculate new cluster centers *Cn+1*  *Diff<= Euclidian (Cn, Cn+1)*  **while** (*Diff <THRESHOLD*) |

The DryadLINQ implementation uses an Apply operation, which executes in parallel in terms of the data vectors, to calculate the partial cluster centers. Another Apply operation, which runs sequentially, calculates the new cluster centers for the *nth* iteration. Finally, we calculate the distance between the previous cluster centers and the new cluster centers using a Join operation to compute the Euclidian distance between the corresponding cluster centers. DryadLINQ support “loop unrolling”, using which multiple iterations of the computation can be performed as a single DryadLINQ query. Deferred query evaluation is a feature of LINQ, whereby a query is not evaluated until the program accesses the query results.. Thus, in the K-means program, we accumulate the computations performed in several iterations (we used 4 as our unrolling factor) into one query and only “materialize” the value of the new cluster centers every 4th iteration. In Hadoop’s MapReduce model, each iteration is represented as a separate MapReduce computation. Notice that without the loop unrolling feature in DryadLINQ, each iteration would be represented by a separate execution graph as well.

* + 1. Evaluations and Findings

When implementing K-means algorithm using DryadLinq we noticed that the trivial MapReduce style implementation of this algorithm perform extremely slow. We had to make several optimizations to the data structures and how we perform the calculations. One of the key changes is the use of Apply operation instead of Select to compare each data point with the current set of cluster centers. This enables DryadLinq to consume an entire data partition at once and perform the comparisons. Figure 12 shows a comparison of performances of different implementations of K-means clustering.

****

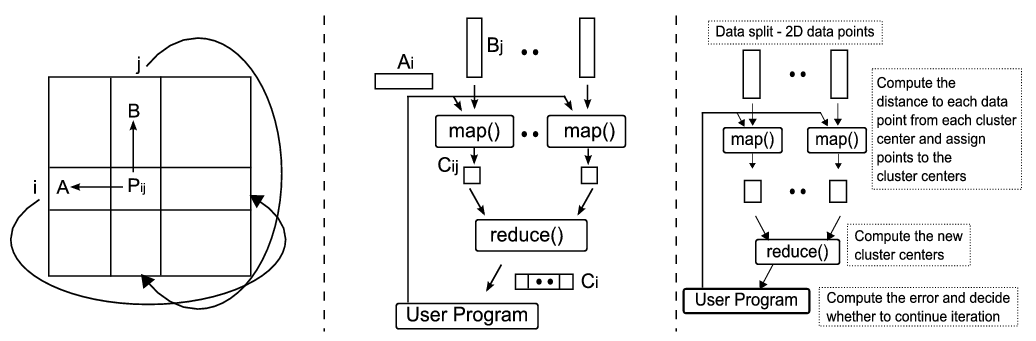
1. Performance of different implementations of clustering algorithm.

The performance graph shows that although DryadLinq performs better than Hadoop for K-means application, still the average time taken by DryadLinq and Hadoop implementations is extremely large compared to the MPI and the CGL-MapReduce implementations.

Although we used a fixed number of iterations, we changed the number of data points from 500k to 20 millions. Increase in the number of data points triggers the amount of computation. However, it was not sufficient to ameliorate the overheads introduced by Hadoop and DryadLINQ runtimes. As a result, the graph in Figure 12 mainly shows the overhead of the different runtimes. With its loop unrolling feature, DryadLinq does not need to materialize the outputs of the queries used in the program in every iteration. In the Hadoop implementation each iteration produces a new MapReduce computation increasing the total overhead of the implementation. The use of file system based communication mechanisms and the loading of static input data at each iteration (in Hadoop) and in each unrolled loop (in DryadLINQ) results in higher overheads compared to CGL-MapReduce and MPI. Iterative applications which perform more computations or access larger volumes of data may produce better results for Hadoop and DryadLINQ as the higher overhead induced by these runtimes becomes relatively less significant. Currently the academic release uses file system based communication mechanism. However, according to the architecture discussed in Dryad paper , Dryad is capable of communicating via TCP pipes and therefore we expect better performances for this type of applications once it is supported by DryadLinq as well.

* + 1. Another Relevant Application - Matrix Multiplication

Parallel applications that are implemented using message passing runtimes can utilize various communication constructs to build diverse communication topologies. For example, a matrix multiplication application that implements Fox's Algorithm and Cannon’s Algorithm assumes parallel processes to be in a rectangular grid. Each parallel process in the grid communicates with its left and top neighbors as shown in figure 13 (left). The current cloud runtimes, which are based on data flow models such as MapReduce and Dryad, do not support this behavior, in which the peer nodes communicate with each other. Therefore, implementing the above type of parallel applications using MapReduce or Dryad requires adopting different algorithms.



1. (Left) The communication topology of Cannon’s Algorithm implemented using MPI, (middle) Communication topology of matrix multiplication application based on MapReduce, and (right) Communication topology of K-means Clustering implemented as a MapReduce application.

We have implemented matrix multiplication applications using Hadoop and CGL-MapReduce by adopting a 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 a set of column blocks and the matrix A into a set of row blocks. In each iteration, all the map tasks process two inputs: (i) a column block of matrix B, and (ii) a row block of matrix A; 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 (a typical MapReduce model), there is no way to specify this behavior. 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 tasks are allowed to retain static data in the memory across invocations, yielding better performance for “Iterative MapReduce” computations. The communication pattern of this application is shown in figure 13 (middle). We haven’t implemented a Matrix multiplication application using DryadLinq yet and plan to do so in the future.

1. Analysis
   1. DryadLinq vs. Other Runtimes
      1. Handling Data

Cloud technologies adopts a more data centered approach to parallel programming compared to the traditional parallel runtimes such as MPI, Workflow runtimes, and individual job scheduling runtimes in which the scheduling decisions are made mainly by the availability of the computation resources. DryadLinq starts its computation from a partition table adapting the same data centered approach and try to schedule computations where the data is available.

In DryadLinq the data is partitioned to the shared directories of the computation nodes of the HPC cluster where all the nodes have access to these common directories. With the support from a *partitioned file* DryadLinq builds the necessary meta-data to access these data partitions and it also supports replicated data partitions to improve the fault tolerance. As we have discussed under sections 3.1 and 3.3.1 with the current release of DryadLinq the partitioning of the existing data (either in individual files or in large data items) needs to be handled by the user manually. Comparatively, Apache Hadoop comes with a distributed file system that can be deployed on top of a set of heterogeneous resources, and a set of client tools to perform necessary file system operations. With this the user is completely shielded from the locations where the data is stored and its fault tolerance functionalities. CGL-MapReduce also adopts a DryadLinq style meta-data model to handle data partitions and currently supports file based data types.

Although the use of a distributed file system in Hadoop makes the data partitioning and managing much easier, not all the applications benefit from this approach. For example, in HEP data analysis, the data is processed via a specialized software framework named ROOT which needs to access data files directly from the file system, but Hadoop provides only Java and C# API to access HDFS. We used a shared parallel file system (Lustre) deployed at Indiana University to store HEP data and this resulted higher overheads in the Hadoop implementation. Apache subprojects such as FUSE allows HDFS to be mounted as a shared file system but we are not sure how the paradigm of moving computation to data works in that approach.

Sector/Sphere is a parallel runtime developed by Y. Gu, and R. L. Grossman that can be used to implement MapReduce style applications. Sphere uses Sector distributed files system resembling an architecture similar to Hadoop.

* + 1. Parallel Topologies

Parallel topologies supported by various parallel runtimes and the problems that can be implemented using these parallel topologies determine the applicability of many parallel runtimes to the problems in hand. For example, many job scheduling infrastructures such as TORQUE and SWARM can be used to execute parallel applications such as CAP3 consisting of a simple parallel topology of a collection of large number of independent tasks. Applications that perform parametric sweeps, document conversions, and brute-force searches are few other examples of this category. DryadLinq, Hadoop, and CGL-MapReduce can all handle this class of applications well. Except for the manual data partitioning requirement, programming such problems using DryadLinq is considerably easier than Hadoop or CGL-MapReduce implementations. With the debugging support from visual studio and the automatic deployment mechanism, the users can develop applications faster with DryadLinq. The CAP3 program we developed using DryadLinq can be used as a model for many similar problems which has the simple parallel topology of collection of independent tasks.

MapReduce programming model provides more parallel topologies than the simple independent tasks with its support for the “reduction” phase. In typical MapReduce model, the outputs of the map tasks are partitioned using a hash function and assigned to a collection of reduce tasks. With the support of overloaded “key selectors” or hashes and by selecting the appropriate key selector function, this simple process can be extended to support additional models producing customized topologies under the umbrella of MapReduce model. For example, in the MapReduce version of tera-sort [16] application, Hadoop uses a customized hashing function to model the bucket sort algorithm. In DryadLinq we can use the programming flows of Apply -> GroupBy -> Apply or Select -> GrouBy -> Apply to simulate MapReduce style computations by using an appropriate GroupBy function.

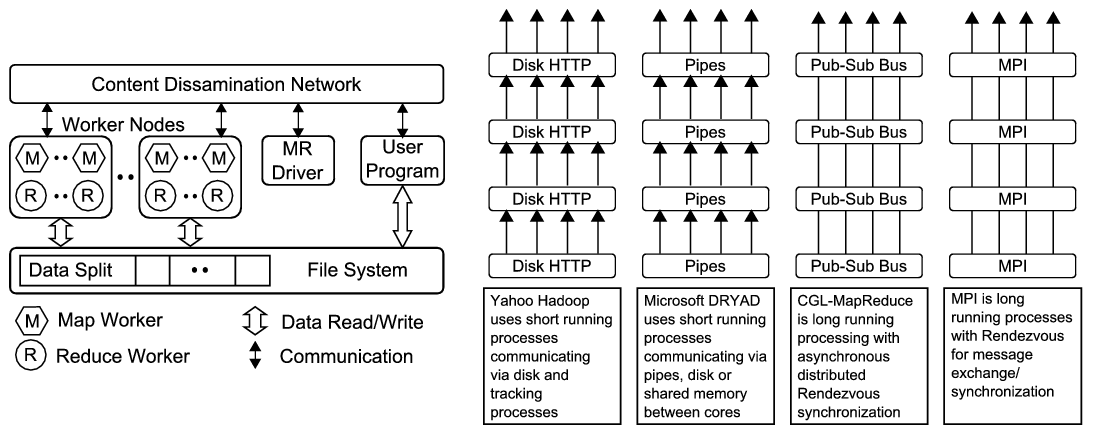
Among other parallel runtimes that support individual tasks and MapReduce style applications, Sphere adopts a streaming based computation model used in GPUs which can be used to develop applications with parallel topologies as a collection of MapReduce style applications. All Pairs solves the specific problems of comparing elements in two data sets with each other and several other specific parallel topologies. We have used DryadLINQ to perform a similar computation to calculate pair-wise distances of a large collection of genes and our algorithm is explained in details in section 3.2. Swift provides a scripting language and a execution and management runtime for developing parallel applications with the added support for defining typed data products via schemas. DryadLINQ allows user to define data types as C# structures or classes allowing users to handle various data types seamlessly with the runtime with the advantage of strong typing. Hadoop allows user to define “record readers” depending on the data that needs to be processed.

Parallel runtimes that support DAG based execution flows provide more parallel topologies compared to the mere MapReduce programming model or the models that support scheduling of large number of individual jobs. Condor DAGMan is a well-known parallel runtime that supports applications expressible as DAGs and many workflow runtimes supports DAG based execution flows. However, the granularity of tasks handled at the vertices of Dryad/DryadLINQ and the tasks handled at map/reduce tasks in MapReduce is more fine grained than the tasks handled in Condor DAGMan and other workflow runtimes. This distinction become blurred when it comes to the parallel applications such as CAP3 where the entire application can be viewed as a collection of independent jobs, but for many other applications the parallel tasks of cloud technologies such as Hadoop and Dryad are more fine grained than the ones in workflow runtimes. For example, during the processing of the *GroupBy* operation used in DryadLINQ, which can be used to group a collection of records using a user defined key field, a vertex of the DAG generated for this operation may only process few records. In contrary the vertices DAGMan may be a complete programs performing considerable amount of processing.

Although in our analysis we compared DryadLINQ with Hadoop, DryadLINQ provides higher level language support for data processing than Hadoop. Hadoop’s sub project Pig is a more natural comparison to DryadLINQ. Our experience suggests that the scientific applications we used maps more naturally to Hadoop and Dryad (currently not available for public use) programming models than the high level runtimes such as Pig and DryadLINQ. However, we expect the high level programming models provided by the runtimes such as DryadLINQ and Pig are more suitable for applications that process structured data that can be fit into tabular structures.

* + 1. MapReduce++

Our work on CGL-MapReduce (we called it MapReduce++) extends capabilities of the MapReduce programming to applications that perform iterative MapReduce computations. We differentiate the variable and fixed data items used in MapReduce computation and allow cacheable map/reduce tasks to hold static data in memory to support faster iterative MapReduce computations. The use of streaming for communication enables MapReduce++ to operate with minimum overheads. Currently CGL-MapReduce does not provide any fault tolerance support for applications and we are investigating the mechanisms to support fault tolerance with the streaming based communication mechanisms we use. The architecture of CGL-MapReduce and a comparison of synchronization and intercommunication mechanisms used by the parallel runtimes are shown in figure 14.



1. (Left) Components of the CGL-MapReduce. (Right) Different synchronization and intercommunication mechanisms used by the parallel runtimes.
   1. Performance and Usability of Dryad

We have applied DryadLINQ to a series of data/compute intensive applications with unique requirements. The applications range from simple map-only operations such as CAP3 MapReduce jobs in HEP data analysis and iterative MapReduce in K-means clustering. We showed that all these applications can be implemented using the DAG based programming model of DryadLINQ, and their performances are comparable to the MapReduce implementations of the same applications developed using Hadoop.

We also observed that cloud technologies such as DryadLINQ and Hadoop work well for many applications with simple communication topologies. The rich set of programming constructs available in DryadLINQ allows the users to develop such applications with minimum programming effort. However, we noticed that higher level of abstractions in DryadLINQ model sometimes make fine-tuning the applications more challenging.

Hadoop and DryadLINQ differ in their approach to fully utilize the many cores available on today’s compute nodes. Hadoop allows scheduling of a worker process per core. On the other hand, DryadLINQ assigns vertices (i.e. worker processes) to nodes and achieves multi-core parallelism with PLINQ. The simplicity and flexibility of the Hadoop model proved effective for some of our benchmarks. The coarser granularity of scheduling offered by DryadLINQ performed equally well once we got a version DryadLINQ working with a newer build of the PLINQ library. Future releases of DryadLINQ and PLINQ will make those improvements available to the wider community. They will remove current needs for manual fine-tuning, which could also be alleviated by adding a tuning option that would allow a DryadLINQ user to choose the scheduling mode that best fits their workload.

Features such as loop unrolling let DryadLINQ perform iterative applications faster, but still the amount of overheads in DryadLINQ and Hadoop is extremely large for this type of applications compared to other runtimes such as MPI and CGL-MapReduce.

Apart from those we would like to highlight the following usability characteristics of DraydLinq comparing it with other similar runtimes.

* + 1. Installation and Cluster Access

We note a technical issue we encountered using Dryad within our Windows HPC environment.  The HPC clusters at our institution are setup using a network configuration that has the headnode connected directly to the enterprise network (ADS domain access) and the compute nodes behind the headnode on a private network.  Enterprise network access is provided to the compute nodes via DHCP and NAT(network address translation) services running on the headnode.  This is our preferred configuration as it isolates the compute nodes from extraneous network traffic, places the compute nodes on a more secure private network and minimizes the attack surface of our clusters.

Using this configuration with Dryad has been somewhat cumbersome as the configuration does not allow direct access to the compute node’s private network from the enterprise network.  In other words, unless we run our dryad jobs directly on the headnode we are unable to access the compute node file systems as only the headnode is aware of the private network.

* + 1. Developing and Deployment of Applications

Enabling DryadLinq for an application simply requires adding DryadLinq.dll to the project and pointing to the correct DraydLinqConfig.xml. After this step, the user can develop applications using Visual Studio and use it to deploy and run DryadLinq applications directly on the cluster. With the appropriate cluster configurations, the development teams can test DryadLinq applications directly from their workstations. In Hadoop, the user can add Hadoop jar files to the class path and start developing Hadoop applications using a Java development environment, but to deploy and run those applications the user need to create jar files packaging all the necessary programs and then copy them to a particular directory that Hadoop can find. Tools such as IBM’s eclipse plugin for MapReduce add more flexibility to create MapReduce computations using Hadoop.

* + 1. Debugging

DryadLinq supports debugging applications via visual studio by setting the property DraydLinq.LocalDebug=true. This is a significant improvement of usability compared to the other parallel runtimes such as Hadoop. The user can simply develop the entire application logic in his workstation and move to the cluster to do the actual data processing. Hadoop also supports single machine deployments but the user needs to do manual configuration and debugging to test applications.

* + 1. Monitoring

DryadLinq depends on the HCP Cluster Manager and HPC Job Manager’s monitoring capabilities to monitor the progress and problems of the jobs. Although the HPC Cluster Manager and Job Manager give better view of the hardware utilization and locations where the job getting executed, there is no direct way to find the progress of the DryadLinq applications. Finding an error that happens only in a cluster deployment is even harder with the current release of DraydLinq. For example, the user need to follow the steps below to find the standard output (stdout) and standard error (stderr) streams related to a particular vertex of the DraydLinq application.

1. Find the job’s ID using Job Manager
2. Find which vertex (sub job has failed) and find its task number
3. Find where that task was running using Job Manager
4. Navigate to the shared directory where the job outputs are created
5. Open the stdout and stderr files to find any problems.

Note: When the vertex is using an Apply operation even that approach won’t give any information because then the standard outputs printed by the program does not get saved in stdout or stderr files.

Hadoop provides a simple web interface to monitor the progress of the computations and to locate these standard output and error files. A simple view of how many map/reduce tasks completed so far gives a better understanding of the progress of the program in Hadoop. We think that a simple approach like this would help new users to develop applications easily without frustration using DryadLinq.

1. Summary of key features of applications that suitable/not for Dryad

In the past Fox has discussed the mapping applications to different hardware and software in terms of 5 “Application Architectures” . These 5 categories are listed in Table 3.

1. Application classification

|  |  |  |
| --- | --- | --- |
| 1 | Synchronous | The problem can be implemented with instruction level Lockstep Operation as in SIMD architectures |
| 2 | Loosely Synchronous | These problems exhibit iterative Compute-Communication stages with independent compute (map) operations for each CPU that are synchronized with a communication step. This problem class covers many successful MPI applications including partial differential equation solution and particle dynamics applications. |
| 3 | Asynchronous | Compute Chess and Integer Programming; Combinatorial Search often supported by dynamic threads. This is rarely important in scientific computing but at heart of operating systems and concurrency in consumer applications such as Microsoft Word. |
| 4 | Pleasingly Parallel | Each component is independent. In 1988, Fox estimated this at 20% of the total number of applications but that percentage has grown with the use of Grids and data analysis applications as seen here and for example in the LHC analysis for particle physics . |
| 5 | Metaproblems | These are coarse grain (asynchronous or dataflow) combinations of classes 1)-4). This area has also grown in importance and is well supported by Grids and described by workflow. |
| 6 | MapReduce++ | It describes file(database) to file(database) operations which has three subcategories given below and in figure 5.   * 1. Pleasingly Parallel Map Only   2. Map followed by reductions   3. Iterative “Map followed by reductions” – Extension of Current Technologies that supports much linear algebra and datamining |

The above classification 1 to 5 largely described simulations and was not aimed directly at data processing. Now we can use the introduction of MapReduce as a new class which subsumes aspects of classes 2, 4, 5 above. We generalize MapReduce to include iterative computations and term it MapReduce++. We have developed a prototype of this extended model and term it currently CGL-MapReduce [8]. Then this new category is summarized as:

Note overheads in categories 1, 2, 6c go like Communication Time/Calculation Time and basic MapReduce pays file read/write costs while MPI overhead is measured in microseconds. In CGL-MapReduce we use data streaming to reduce overheads while retaining the flexibility and fault-tolerance of MapReduce. MapReduce++ supports the Broadcast and Reduce operations in MPI which are all that is needed for much linear algebra and datamining including the clustering and MDS approaches described earlier.

1. Comparison of MapReduce++ subcategories and Loosely Synchronous category

|  |  |  |  |
| --- | --- | --- | --- |
| **Map-only** | **Classic**  **Map-reduce** | **Iterative Reductions**  **MapReduce++** | **Loosely**  **Synchronous** |
| D:\academic\phd\Publications\eScience2009\gnuplot\map-only.png | D:\academic\phd\Publications\eScience2009\gnuplot\map-reduce.png | D:\academic\phd\Publications\eScience2009\gnuplot\map-reduce-iter.png | D:\academic\phd\Publications\CloudComp09\figures\loosely-sync-MPI.png |
| * Document conversion   (PDF->HTML)   * Brute force searches in cryptography * Parametric sweeps * CAP3 Gene assembly * PolarGrid Matlab data analysis | * High Energy Physics   (HEP) Histograms   * Distributed search * Distributed sort * Information retrieval * Calculation of Pairwise Distances for ALU sequences | * Expectation maximization algorithms * Linear Algebra * Datamining including * Clustering * K-means * Deterministic   Annealing clustering   * Multidimensional Scaling (MDS) | * Many MPI scientific applications utilizing wide variety of communication constructs including local interactions * Solving differential equations and * Particle dynamics with short range forces |
| **Domain of MapReduce and Iterative Extensions** | | | **MPI** |

From the applications we developed it is trivial that the DryadLinq can be applied to real scientific analyses. DryadLinq performs competitively well with Hadoop for both pleasingly parallel and MapReduce style applications. However, applicability of DryadLinq (also Hadoop) for iterative MapReduce applications is questionable. The file based communication mechanism and loading of static data again and again causes higher overheads in this class of applications. However, we expect that these overheads may reduce if DryadLinq support in memory communication mechanism such as TCP pipes.

The scheduling inefficiencies present in the current PLINQ library hinder the performance of DryadLinq applications. The user may need to perform additional optimization so achieve better CPU utilization. However, we have verified that the latest build of PLINQ has solved these inefficiencies and DryadLinq perform as expected when it is used with the new build of PLINQ.

Additional support for partitioning data (few tools to perform various data partitioning strategies) and a mechanism to monitoring the progress of applications are two areas that DryadLinq needs improvements.

References

1. X. Huang and A. Madan, “CAP3: A DNA Sequence Assembly Program,” Genome Research, vol. 9, no. 9, pp. 868-877, 1999.
2. J. Hartigan. Clustering Algorithms. Wiley, 1975.
3. ROOT Data Analysis Framework, <http://root.cern.ch/drupal/>
4. 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.
5. Y.Yu, M. Isard, D. Fetterly, M. Budiu, Ú. Erlingsson, P. Gunda, 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), CA, December 8-10, 2008.
6. Apache Hadoop, http://hadoop.apache.org/core/
7. J. Dean, and S. Ghemawat. 2008. MapReduce: simplified data processing on large clusters. *Commun. ACM* **51**(1): 107-113.
8. J. Ekanayake, S. Pallickara, and G. Fox, “MapReduce for Data Intensive Scientific Analysis,” Fourth IEEE International Conference on eScience, 2008, pp.277-284.
9. Geoffrey Fox, Seung-Hee Bae, Jaliya Ekanayake, Xiaohong Qiu, and Huapeng Yuan: Parallel Data Mining from Multicore to Cloudy Grids. Proceedings of HPC 2008 High Performance Computing and Grids workshop Cetraro Italy July 3 2008
10. MPI (Message Passing Interface), <http://www-unix.mcs.anl.gov/mpi/>
11. J. B. Kruskal and M.Wish. *Multidimensional Scaling*. Sage Publications Inc., Beverly Hills, CA, U.S.A., 1978.
12. Fox, G. C., Hey, A. and Otto, S., Matrix Algorithms on the Hypercube I: Matrix Multiplication, *Parallel Computing*, 4, 17 (1987)
13. Johnsson, S. L., T. Harris, et al. 1989. Matrix multiplication on the connection machine. Proc of the *1989 ACM/IEEE conference on Supercomputing*. Reno, Nevada, United States, ACM.
14. Mountable HDFS, http://wiki.apache.org/hadoop/MountableHDFS
15. Y. Gu, and R. L. Grossman. 2009. Sector and Sphere: the design and implementation of a high-performance data cloud. *Philos Transact A Math Phys Eng Sci* **367**(1897): 2429-45.
16. Torque Resource Manager, http://www.clusterresources.com/products/torque-resource-manager.php
17. S. Pallickara, and M. Pierce. 2008. SWARM: Scheduling Large-Scale Jobs over the Loosely-Coupled HPC Clusters. Proc of *IEEE Fourth International Conference on eScience '08(eScience, 2008)*.Indianapolis, USA
18. C. Moretti, H. Bui, K. Hollingsworth, B. Rich, P. Flynn, D. Thain, "All-Pairs: An Abstraction for Data Intensive Computing on Campus Grids," IEEE Transactions on Parallel and Distributed Systems, 13 Mar. 2009.
19. Zhao Y., Hategan, M., Clifford, B., Foster, I., vonLaszewski, G., Raicu, I., Stef-Praun, T. and Wilde, M Swift: Fast, Reliable, Loosely Coupled Parallel Computation IEEE International Workshop on Scientific Workflows 2007
20. Codor DAGMan, http://www.cs.wisc.edu/condor/dagman/.
21. Apache Pig project, http://hadoop.apache.org/pig/
22. Geoffrey C. Fox, Roy D. Williams, Paul C. Messina, *Parallel Computing Works!* Morgan Kaufmann (1994).
23. Enabling Grids for E-science (EGEE): <http://www.eu-egee.org/>
24. IBM Eclipse plugin for MapReduce, <http://www.alphaworks.ibm.com/tech/mapreducetools>
25. J. Ekanayake, A. S. Balkir, T. Gunarathne, G. Fox, C. Poulain, N. Araujo, R. Barga. "DryadLINQ for Scientific Analyses", Technical report, Accepted for publication in eScience 2009
26. M.A. Batzer, P.L. Deininger, 2002. "Alu Repeats And Human Genomic Diversity." Nature Reviews Genetics 3, no. 5: 370-379. 2002
27. A. F. A. Smit, R. Hubley, P. Green, 2004. Repeatmasker. http://www.repeatmasker.org
28. J. Jurka, 2000. Repbase Update: a database and an electronic journal of repetitive elements. Trends Genet. 9:418-420 (2000).
29. Source Code. Smith Waterman Software. http://jaligner.sourceforge.net
30. T.F. Smith, M.S.Waterman,. Identification of common molecular subsequences. Journal of Molecular Biology 147:195-197, 1981.
31. O. Gotoh, An improved algorithm for matching biological sequences. Journal of Molecular Biology 162:705-708 1982.