**Using MapReduce Technologies in Bioinformatics and Medical Informatics**

Xiaohong Qiu1, Jaliya Ekanayake 1,2, Thilina Gunarathne1,2, Seung-Hee Bae1,2,Jong Youl Choi1,2, Scott Beason1, Geoffrey Fox1,2

1Pervasive Technology Institute, 2School of Informatics and Computing,

Indiana University

Bloomington IN, U.S.A.

{ xqiu, jekanaya, tgunarat, sebae, jychoi, smbeason, gcf@indiana.edu}

There have been several important commercial developments of computing technologies that have important implications for scientific computing. Cloud computing is best known for the systems like Amazon EC2, Eucalyptus and Azure which use virtual machines to provide flexible, dynamic, easy to use computing on demand. Another important development is MapReduce systems that were developed to support the huge information retrieval industry. This is perhaps the largest data analysis problem and so it is particularly interesting to examine for scientific data processing which is of growing importance as the data deluge continues. We have examined MapReduce for several applications including particle physics and several biology and medical informatics cases. We have looked at both Hadoop (Yahoo) and Dryad (Microsoft) and compared them seeing similar performance and here we focus on Dryad where we have recently completed studies on our 768 core Windows HPC Server cluster Tempest [1-5]. Four applications we have looked at in detail are:

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

b) Pairwise Alu gene alignment using Smith Waterman dissimilarity computations followed by MPI applications for Clustering and MDS (Multi Dimensional Scaling)

c) Correlating Childhood obesity with environmental factors by combining medical records with Geographical Information data with over 100 attributes using correlation computation, MDS and genetic algorithms for choosing optimal environmental factors.

d) Mapping the 26 million entries in PubChem into two or three dimensions to aid selection of related chemicals with convenient Google Earth like Browser. This uses either hierarchical MDS (which cannot be applied directly as O(N2)) or GTM (Generative Topographic Mapping).

These applications have common and individually distinctive patterns. All have data parallel steps that can directly use MapReduce and these steps are a significant part of computation and for (a) and (d) (MDS version) dominant. These MapReduce steps are usually “Doubly Data Parallel” with independent parallelism over two datasets that are sometimes identical. Further application (a) is very heterogeneous with individual computations varying drastically in compute time. The others have approximately uniform computational complexities for each computation and these can be easily load balanced statistically. More research is needed on support of heterogeneous datasets in MapReduce. We sometimes need to combine the natural MapReduce steps with following data mining applications (such as MDS, GTM, and Clustering) that must use parallelism and for which MPI is suitable. The current Hadoop and Dryad have poor performance if used these applications although MPI can be programmed only to use reductions for these cases. MPI efficiently supports iterative “Map” followed by “Reduce” keeping information in memory rather than file systems. We have developed CGL-MapReduce which is a version of MapReduce that supports such iterative applications and compared its performance with MPI. It has higher overheads but for large enough problems it gets excellent parallel performance. It is not clear if the natural model is MapReduce followed by MPI or a single environment supporting both. We used not just the basic operations in MapReduce but also operations such as the “homomorphic Apply” in Dryad. In the cases with follow-on MPI steps, we showed that Dryad can be programmed to prepare data for use in later data-mining applications. This involved generating a matrix from the doubly data parallel initial step and this could be a rather general programming pattern. The languages that drive MapReduce have some similarities with workflow and one can wonder whether integrated environments would support workflow, MapReduce (file parallelism) and MPI (memory parallelism). We believe that enhanced MapReduce can support a broad range of systems biology application with performance competitive with MPI but with greater flexibility and fault tolerance. Exactly which enhancements should be put into MapReduce and which should be separate but linked needs further research. Heterogeneous datasets also have many open issues.

[1] 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 <http://grids.ucs.indiana.edu/ptliupages/publications/CetraroWriteupJune11-09.pdf>

[2] Jaliya Ekanayake, Geoffrey Fox “High Performance Parallel Computing with Clouds and Cloud Technologies”, First International Conference CloudComp on Cloud Computing October 19 - 21, 2009, Munich, Germany <http://grids.ucs.indiana.edu/ptliupages/publications/cloudcomp_camera_ready.pdf>

[3] Geoffrey Fox, Xiaohong Qiu, Scott Beason, Jong Youl Choi, Mina Rho, Haixu Tang, Neil Devadasan, Gilbert Liu “Case Studies in Data Intensive Computing: Large Scale DNA Sequence Analysis as the Million Sequence Challenge and Biomedical Computing” Technical Report 9 August 2009 <http://grids.ucs.indiana.edu/ptliupages/publications/UsesCasesforDIC-Aug%209-09.pdf>

[4] Jaliya Ekanayake, Xiaohong Qiu, Thilina Gunarathne, Scott Beason, Geoffrey Fox “High Performance Parallel Computing with Clouds and Cloud Technologies” August 25 2009 to appear as Book Chapter <http://grids.ucs.indiana.edu/ptliupages/publications/cloud_handbook_final-with-diagrams.pdf>

[5] Xiaohong Qiu, Jaliya Ekanayake, Scott Beason, Thilina Gunarathne, Geoffrey Fox, Roger Barga, Dennis Gannon “Cloud Technologies for Bioinformatics Applications” Technical Report September 8 2009 <http://grids.ucs.indiana.edu/ptliupages/publications/MTAGS09-23.pdf>