**Architectures for Iterative Data Intensive Analysis Computations on Clouds and Heterogeneous Environments**

**Student: Thilina Gunarathne {tgunarat@cs.indiana.edu}**

**Advisor: Prof. Geoffrey Fox {gcf@indiana.edu}**

**School of Informatics and Computing**

**Indiana University, Bloomington, IN 47405**

Iterative computations are at the core of the vast majority of scientific computations. Many important data intensive iterative scientific computations can be implemented as iterative computation and communication steps, where computations inside an iteration are independent and are synchronized at the end of each iteration through reduce and communication steps, making it possible for individual iterations to be parallelized using technologies such as MapReduce. Examples of such applications include dimensional scaling, many clustering algorithms, many machine learning algorithms, expectation maximization applications and many more. The growth of such data intensive iterative computations, in number as well as importance, is driven partly by the need to process massive amounts of data and the emergence of data intensive computational fields, such as bioinformatics, chemical informatics and web mining.

The utility computing model introduced by cloud computing combined with the rich set of cloud infrastructure services offers a very viable environment for the scientists to process massive amounts of data. However, clouds offer unique reliability and sustained performance challenges to large scale computations due to the virtualization, multi-tenancy, non-dedicated commodity connectivity and etc, necessitating distributed parallel computing frameworks specifically tailored for cloud characteristics to harness the power of clouds both easily and effectively. Another important development we note is the emergence of heterogeneous clouds, where clouds offer many-core resources such as GPGPU’s alongside CPU compute resources, making it possible to utilize GPGPU’s for certain computations.. The goal of my PhD research is to identify and develop distributed parallel computation frameworks to facilitate the optimized efficient execution of iterative as well as non-iterative data-intensive computations in cloud environments, along with the evaluation of heterogeneous cloud resources offerings for data-intensive iterative computations. The current issues and challenges include identifying suitable programing abstractions & primitives, performing appropriate task partitioning & task scheduling, identifying suitable data storage architectures, optimizing the performance through multi-level data reuse/caching for iterative computations, identifying suitable communication patterns for intermediate data communication & data broadcasting and exploring appropriate fault tolerance mechanisms.

**Twister4Azure** is a distributed decentralized **iterative** MapReduce runtime for Windows Azure Cloud that was developed utilizing Azure cloud infrastructure services as part of pursuing the goal of my PhD research. Twister4Azure extends the familiar, easy-to-use MapReduce programming model with iterative extensions, enabling a wide array of large scale iterative as well as non-iterative data analysis and scientific applications to utilize Azure platform easily and efficiently in a fault-tolerant manner. Twister4Azure utilize the eventually-consistent, high-latency Azure cloud services effectively to deliver performance comparable to (non-iterative) and outperforming (for iterative computing) traditional MapReduce runtimes. Twister4Azure has minimal management & maintenance overheads and provides users with the capability to dynamically scale up or down the amount of compute resources. I’m also evaluating the feasibility of GPGPUs for this class of applications in determining the potential of combining GPGPU computing together with iterative MapReduce frameworks, to utilize the heterogeneous cloud resources effectively.

More information at <https://www.cs.indiana.edu/~tgunarat/> and <http://salsahpc.indiana.edu/twister4azure/>

**Selected Publications**

[1] **T. Gunarathne**, T. L. Wu, J. Qiu, and G. C. Fox, "MapReduce in the Clouds for Science," presented at the 2nd International Conference on Cloud Computing, Indianapolis, 2010.

[2] **T. Gunarathne**, J. Qui, and G. Fox, "Iterative MapReduce for Azure Cloud," presented at the Cloud Computing and Its Applications, ANL, Chicago, IL, 2011.

[3] **T. Gunarathne**, T.-L. Wu, J. Y. Choi, S.-H. Bae, and J. Qiu, "Cloud computing paradigms for pleasingly parallel biomedical applications," *Concurrency and Computation: Practice and Experience,* 2011.

[4] **T. Gunarathne**, T.-L. Wu, J. Qiu, and G. Fox, "Cloud Computing Paradigms for Pleasingly Parallel Biomedical Applications," presented at the Proceedings of the Emerging Computational Methods for the Life Sciences Workshop of ACM HPDC 2010 conference, Chicago, Illinois, 2010.

[5] J. Ekanayake, **T. Gunarathne**, and J. Qiu, "Cloud Technologies for Bioinformatics Applications," *Parallel and Distributed Systems, IEEE Transactions on,* vol. 22, pp. 998-1011, 2011.

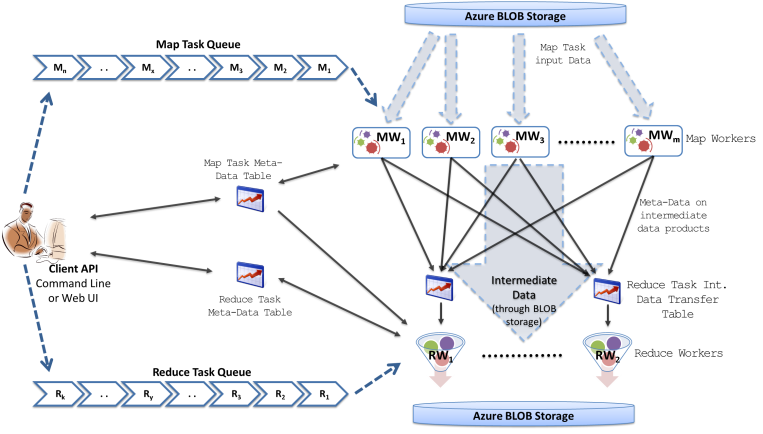
[6] J.Ekanayake, H.Li, B.Zhang, **T.Gunarathne**, S.Bae, J.Qiu, and G.Fox., "Twister: A Runtime for iterative MapReduce," presented at the Proceedings of the First International Workshop on MapReduce and its Applications of ACM HPDC 2010 conference June 20-25, 2010, Chicago, Illinois, 2010.

[7] J. Ekanayake, A. S. Balkir, **T. Gunarathne**, G. Fox, C. Poulain, N. Araujo, and R. Barga, "DryadLINQ for Scientific Analyses," in *Fifth IEEE International Conference on eScience: 2009*, Oxford, 2009.

**Selected Publications Under Review**

[8] **T. Gunarathne**, B. Salpitikorala, G. C. Fox, and A. Chauhan. Optimizing OpenCL Kernels for Iterative Statistical Applications on GPUs. *Submitted to 2nd International Workshop on GPUs and Scientific Applications, 2011*. Available: https://www.cs.indiana.edu/~tgunarat/gpusca.pdf

[9] **T. Gunarathne**, T.-L. Wu, B. Zhang, J. Qiu, and G. C. Fox. (2011, Scientific Applications of Twister4Azure. *Submitted to IEEE International Conference on Utility and Cloud Computing (UCC 2011)*. Available: https://www.cs.indiana.edu/~tgunarat/twister4azure-ucc.pdf

Twister4Azure programming model

Twister4Azure high-level architecture