Name: Judy Qiu

Organization: Indiana University, School of Informatics and Computing

Email: xqiu@indiana.edu

2 Topics from list

4. Programming Models for the Cloud
11. Cloud Test-Beds

Current Cloud Research

My cloud research activity has central theme of looking at programming models with a strong focus on their applicability to scientific computing – especially in the life science area. This work has often used MapReduce with performance and functionality evaluations of both Hadoop and Dryad. I have identified the importance of Iterative MapReduce where I have designed and released a prototype of this called Twister http://www.iterativemapreduce.org/ with student Jaliya Ekanayake. In Ekanayake's PhD thesis and two papers, we have explored the functionality and performance of Twister showing it can reproduce previous successes of Dryad and Hadoop but also support important data mining algorithms including Page Rank, Latent Dirichlet Allocation, Clustering and Dimension Reduction. We have also applied it successfully to kernels like matrix multiplication. We are now exploring fault tolerance and communication models that will allow Twister to scale well on a broad class of data intensive problems.

On a broader level, I was PC co-chair of the 2nd IEEE International CloudCom conference in Indianapolis from November 30-December 3 2010 which had 296 registrants and around 140 total papers (25% acceptance in main track), workshops, short papers and posters. This gave me a good overview of the current status of Cloud research. I also have several significant cloud related educational activities http://salsahpc.indiana.edu/tutorial/index.html. We had 200 students in 10 formal class rooms across the country and another 100 online for the Big Data virtual summer school July 2010 which had many cloud presentations and hands-on activities. I offered a Cloud special topics class this fall and my core distributed systems class this semester with 60 students has a significant cloud component this semester. I am building a rich online resource to support this.

Future Research Interests: Iterative MapReduce for Scalable Data Intensive Applications

Cloud computing offers new approaches for scientific computing that leverage the major commercial hardware and software investment in this area. Closely coupled applications are still unclear in clouds as synchronization costs are still higher than on optimized MPI machines. However loosely coupled problems are very important in many fields and can achieve good cloud performance even when pleasingly parallel steps are followed by reduction operations as supported by MapReduce. It appears that many data analysis problems fit the MapReduce paradigm but there is no definitive analysis here. For example analysis of LHC (Large Hadron Collider) data corresponds to a data selection step followed by forming histograms; this naturally corresponds "perfectly" to the MapReduce paradigm. In Life Science, "all-pairs" applications like BLAST can run well with MapReduce but are particularly simple corresponding to "pleasingly parallel" or "map only" structure. Finally there are applications involving steps like the dimension reduction or clustering algorithms (which we have studied in detail) where pleasing parallel operations (such as alignment and sequence distance computation) are followed by data mining steps involving iterative operations – such as those present in matrix algebra. Such iterative algorithms are the mainstay of large scale scientific computing and are linked directly to data with data assimilation in weather and climate area. Even in the "birthplace" of MapReduce – Information Retrieval – the Page Rank algorithm needs iterative MapReduce. Thus we pose the following research questions.

1) What data analysis problems in science can use clouds and/or MapReduce
2) What data analysis problems need iterative algorithms poorly supported by basic MapReduce
3) What are tradeoffs in performance, usability, flexibility and fault tolerance between MPI and Iterative MapReduce
4) What run time can combine the fault tolerance of MapReduce with the performance of MPI on iterative problems.
5) What are requirements for workflow systems needed to support complicated science data processing which uses MapReduce in some of its steps
6) How is the analysis affected by distributed file system used.

We are investigating these questions using the open source Twister environment.

References

1. Jaliya Ekanayake, Hui Li, Bingjing Zhang, Thilina Gunarathne, Seung-Hee Bae, Judy Qiu, Geoffrey Fox Twister: A Runtime for Iterative MapReduce Proceedings of the First International Workshop on MapReduce and its Applications of ACM HPDC 2010 conference, Chicago, Illinois, June 20-25, 2010.
2. Bingjing Zhang, Yang Ruan, Tak-Lon Wu, Judy Qiu, Adam Hughes, Geoffrey Fox Applying Twister to Scientific Applications Proceedings of CloudCom 2010 Conference IUPUI Conference Center Indianapolis November 30-December 3 2010
3. Judy Qiu, Jaliya Ekanayake, Thilina Gunarathne, Jong Youl Choi, Seung-Hee Bae, Hui Li, Bingjing Zhang, Tak-Lon Wu, Yang Ryan, Saliya Ekanayake, Adam Hughes, Geoffrey Fox Hybrid cloud and cluster computing paradigms for life science applications Proceedings of BOSC 2010 published in in BMC Bioinformatics October 6 2010