Name: Geoffrey Fox

Organization: Indiana University, School of Informatics and Computing

Email: gcf@indiana.edu

2 Topics from list

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

Current Cloud Research

I am currently leading FutureGrid which is a testbed for Grid Cloud and HPC including research and education. Interpreting clouds broadly to include technologies such as MapReduce, over half of the 75 current FutureGrid projects are cloud related. As part of FutureGrid we are developing some novel middleware including that supporting dynamic provisioning, security review of user images, experiment management with provenance at all levels from Images up. The dynamic provisioning allows such diverse environments as MPI, OpenMP, Hadoop, Dryad, gLite, Unicore, Globus, Xen, ScaleMP (distributed Shared Memory), Nimbus, Eucalyptus, OpenStack, KVM, and Windows to be investigated on identical hardware. The funded partners in FutureGrid are Indiana University, University of California – San Diego, University of Chicago, University of Florida, University of Texas at Austin – Texas Advanced Computer Center , University of Southern California, University of Tennessee – Knoxville, and University of Virginia with the first 5 institutions supporting hardware.

We have a few cloud research activities with central theme on looking at applications of clouds 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. We have identified the importance of Iterative MapReduce where Judy Qiu and (graduated) student Jaliya Ekanayake have designed and released a prototype Twister of this. We are also starting research into the different “data parallel” file systems like HDFS and Sector to understand how important it is to bring the computation to the data.

Future Research Interests

Future research associated with FutureGrid will fall into two classes

a) Evaluation of emerging technologies for IaaS such as OpenStack and OpenNebula which we will deploy over next 6 months after we make their security models consistent with FutureGrid. OpenStack has a particularly interesting storage solution coming from Rackspace.

b) Improving the core infrastructure especially aiming at ease of use for education uses of FutureGrid that do not require privileged access.

Our broader computer science research will focus on data intensive computing noting that most large scale scientific computing is designed around classic shared file systems whereas the file systems popularized in clouds are "data-parallel" with the data distributed among compute nodes with replication (no back up to archival system) used for both fault tolerance and to allow better of shared systems that do not commit a set of nodes to a particular application and its data. We would like to understand the tradeoff between the traditional HPC hierarchical data architecture and the replicated distributed approach. Correspondingly we need to understand the appropriate general runtime interpolating between MPI-IO and Hbase. We intend to use both conventional MapReduce, Iterative MapReduce and MPI in these environments. We are interested in both "simple" clouds and hybrid systems where for privacy or performance reasons the processing is split between say a public and private cloud. Data parallel approaches to parallel programming such as HPF and the later Darpa HPCS environments (X-10, Chapel, Fortress) seemed promising but so far have not seen wide adoption – partly because it is hard to develop efficient compilers for today's dynamic irregular science simulation problems. However these ideas are sound (indeed Matlab exploits them to some extent) and essentially the same idea can be seen in languages like Sawzall and Pig latin and we are extending these data parallel analysis languages with constructs that take good use of Iterative MapReduce (and transparently other run times) and the various file systems discussed above. Initial results are promising and do not suffer from the problems that HPF had as most data analysis does not have the complex data dependencies of a scientific simulation. We will experiment with these approaches on large scale text processing and life science applications. The former will use new global inferences algorithms exploiting deterministic annealing.