Syllabus for I590 Big Data Open Source Software and Projects Fall 2014

Instructor Geoffrey Fox

This course studies software used in many commercial activities to study Big Data. The backdrop for course is the ~120 software subsystems illustrated at http://hpc-abds.org/kaleidoscope/. We will describe the software architecture represented by this collection which we term HPC-ABDS (High Performance Computing enhanced Apache Big Data Stack). A paper discussing this can be found at http://arxiv.org/abs/1403.1528 and presentations at http://www.slideshare.net/Foxsden/microsoft-april302014 and http://www.slideshare.net/Foxsden/multifaceted-classification-of-big-data-uses-and-proposed-architecture-integrating-high-performance-computing-and-the-apache-stack. Copies of this material may be found at http://www.infomall.org/I590ABDSSoftware/Resources/.

The course covers the following material

a) The cloud computing architecture underlying ABDS and contrast of this with HPC.
b) The software architecture with its different layers at http://hpc-abds.org/kaleidoscope/ covering broad functionality and rationale for each layer.
c) We will give application examples.
d) Then we will go through selected software systems – about 10% of those in the Kaleidoscope which have been already deployed on FutureGrid systems using OpenStack and Chef recipes.
e) Students will chose one other open source member of Kaleidoscope each and deploy as in d).
f) The main activity of the course will be building a significant project using multiple HPC-ABDS subsystems combined with user code and data.
g) Teams of up to 3 students can be formed with corresponding increase in scope in activities e), f).

Grading will be based on participation (10%), ABDS deployment (30%) and Project (60%). The class will interact with postings on a Google community group. The online section will also interact with Google Hangout or equivalent.

We will probably use FutureSystems facilities and cloud computing experience is helpful but essential. Good working experience with Java is required.