Cloud Technologies for Bioinformatics Applications

Xiaohong Qiu1, Jaliya Ekanayake 1,2, Scott Beason1, Thilina Gunarathne1,2, Geoffrey Fox1,2

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

Indiana University

Bloomington IN, U.S.A.

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

Roger Barga3, Dennis Gannon3

3Microsoft Research

Microsoft Corporation

Redmond WA, U.S.A

{Barga, Dennis.Gannon@microsoft.com}

**Abstract**

Executing large number of independent tasks or tasks that perform minimal inter-task communication in parallel is a common requirement in many domains. In this paper, we present our experience in applying two new MapReduce technologies Dryad and Apache Hadoop to two bioinformatics applications. We also compare with traditional MPI and threading. The applications are an EST (Expressed Sequence Tag) sequence assembly program, and a pairwise Alu gene alignment application. We give a detailed MapReduce performance discussion on a cluster booted with Windows HPCS, Linux or Linux on Xen. All the applications start with a “doubly data parallel step” involving independent data chosen from two similar databases. There are different structures for final stages in each application. We investigate sensitivity to data inhomogeneity and measure the overhead of Xen for MapReduce.