INTRODUCTION

One of the challenges of High Performance Computing (HPC) and bio-computing is the user accessibility. At the Cornell University Computational Biology Service Unit we have developed a suit of computational biology applications for HPC (BioHPC) that allows researchers from biological laboratories to submit their jobs to the parallel cluster through an easy-to-use web interface. They don’t need to deal with parallel job submission, queues, clusters – knowing the application, parameters and input is all that is required.

ARCHITECTURE

The system consists of a web server running the interface (ASP.NET C#), Microsoft SQL server (ADO.NET), compute clusters running Microsoft Windows, ftp server and file server. Two local compute cluster schedulers are supported (CCS and HPC Server 2008), remote clusters can be used via JSDL/HPC Profile. HPC Profile / JSDL connection is implemented for integration of geographically dispersed resources.

Users interact with their jobs and data primarily by a web browser and e-mail. Jobs are submitted through our active web pages or web services. Notification e-mails sent to users upon job submission; start, and completion contain links for job progress monitoring, job cancellation and restart, and results retrieval (by http or ftp). Job and data control functions can also be performed via a recently developed web service interface which enables users to build custom clients independent of the web browser. Through web services, BioHPC is being integrated with the Microsoft Biology Foundation platform.

APPLICATIONS AND USAGE

37 applications offered by BioHPC cover various aspects of computational biology: data mining/sequence, protein structure prediction and modeling, population genetics, phylogenetics, association analysis/statistic, MSR Biomedical applications. 140,000 BioHPC jobs have been submitted by 11 471 users from 83 countries, the majority (57% by CPU time used) coming from the USA.

NEXT GENERATION SEQUENCING SUPPORT

Run Manager connects to the sequencing facility, Lane Browser allows users to manage their sequencing read files, Reference Manager allows users to upload and catalogue reference genome files and annotation files needed in downstream data analysis and Pipeline Manager (under development) enables users to construct and run various analysis pipelines using sequencing reads and reference files stored at BioHPC as input.

AVAILABILITY

The interface source code is freely available and can be downloaded from BioHPC.org.

Some applications in BioHPC implementation at Cornell can be used by outside researchers.