INTRODUCTION

One of the challenges of High Performance Computing (HPC) is the user accessibility. At the Cornell University Computational Biology Service Unit, which is also a Microsoft HPC institute, we have developed a suite 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.

TAIR, the major database of the plant model organism Arabidopsis, and SGN, the international tomato genome database, are both using our system for data analysis.

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. JSDL connection is now implemented and linked to Athena cluster at Microsoft (Redmond, WA) and bioins (Cornell).

Users interact with their jobs and data primarily via a web browser and e-mail. Jobs are submitted through our active web pages, fully compatible with all the popular web browsers supporting DOM and Javascript. 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. The number of applications offered through the web service (currently 10) is growing.

Besides the job control interface, BioHPC also features a built-in user and data management system which can limit software and database access to specified users. Data access between different users is restricted – users can access only their own jobs and data. There is also an administrative interface allowing for easy management of jobs, clusters and applications with automatic e-mail notification of possible problems.

APPLICATIONS

Through this system, we are providing users with popular bioinformatics tools covering various aspects of computational biology:

- Data mining / sequence analysis (BLAST, HMMER, InterProScan, GIMSAN, SLIM)
- Protein structure prediction and modeling (LOOPP, Modeller)
- Population genetics (BEAST, BEST, CLUMPP, IM, IMa, InStruct, MDIV, Migrate, MrPRF, MSVAR, Omegamap, Parentage, SFS_CODE, Structurama, Structure, TESS)
- Phylogenetics (MrBayes, ClustalW, Stretchr, T-COFFEE)
- Association analysis / statistics (PLINK, R)
- MSK Biomedical applications (CreateEpitome, Epipred, FalseDiscoveryRate, HisAssignment, HisCompletion, PhyloD)

The system is flexible and can be easily customized to include other software. The interface to each application is standardized, users can choose the cluster, number of nodes or allow the interface to determine it based on the best load balance and node availability. It is also scalable, currently it processes over 40,000 job submissions a year, many of them parallel. BioHPC integrates distributed cluster resources in a user-transparent way.

FUTURE

The interface source code is freely available. The "packaged" version of the interface can be downloaded from BioHPC.net and installed locally with any Microsoft CCS or HPC 2008 cluster.

We are planning to use our web service interface to integrate BioHPC with the Microsoft Biology Foundation. The pool of applications available through the web service will be expanded and a client will be developed as MS Excel add-in. We currently working on support for 2nd generation sequencing pipelines. The tool will be optimized to take better advantage of multi-core nodes.

ABOUT CBSU

Computational Biology Service Unit (CBSU) of the Cornell University Life Sciences Core Laboratories Center was initiated by the Tri-Institutional collaboration among Cornell University, Well Cornell Medical College, Rockefeller University and Memorial Sloan-Kettering Cancer Center. In February 2006 CBSU became Microsoft HPC Institute charter member. CBSU is Cornell core facility for computational biology. BioHPC development is now partially funded by Microsoft Research and CBSU is a Microsoft Biology Initiative partner.

The BioHPC installation at CBSU is currently using 5 Microsoft Windows based local compute clusters totaling 976 cores. The local nodes use Microsoft Server 2003 with CCS and Microsoft Server 2008 with HPC Server 2008. 80 CPU cores of the remote cluster Athena (located in Redmond, WA) are also available via JSDL, courtesy of Microsoft.

INTRODUCTION

One of the challenges of High Performance Computing (HPC) is the user accessibility. At the Cornell University Computational Biology Service Unit, which is also a Microsoft HPC institute, we have developed a suite 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.

TAIR, the major database of the plant model organism Arabidopsis, and SGN, the international tomato genome database, are both using our system for data analysis.

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. JSDL connection is now implemented and linked to Athena cluster at Microsoft (Redmond, WA) and bioins (Cornell).

Users interact with their jobs and data primarily via a web browser and e-mail. Jobs are submitted through our active web pages, fully compatible with all the popular web browsers supporting DOM and Javascript. 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. The number of applications offered through the web service (currently 10) is growing.

Besides the job control interface, BioHPC also features a built-in user and data management system which can limit software and database access to specified users. Data access between different users is restricted – users can access only their own jobs and data. There is also an administrative interface allowing for easy management of jobs, clusters and applications with automatic e-mail notification of possible problems.

APPLICATIONS

Through this system, we are providing users with popular bioinformatics tools covering various aspects of computational biology:

- Data mining / sequence analysis (BLAST, HMMER, InterProScan, GIMSAN, SLIM)
- Protein structure prediction and modeling (LOOPP, Modeller)
- Population genetics (BEAST, BEST, CLUMPP, IM, IMa, InStruct, MDIV, Migrate, MrPRF, MSVAR, Omegamap, Parentage, SFS_CODE, Structurama, Structure, TESS)
- Phylogenetics (MrBayes, ClustalW, Stretchr, T-COFFEE)
- Association analysis / statistics (PLINK, R)
- MSK Biomedical applications (CreateEpitome, Epipred, FalseDiscoveryRate, HisAssignment, HisCompletion, PhyloD)

The system is flexible and can be easily customized to include other software. The interface to each application is standardized, users can choose the cluster, number of nodes or allow the interface to determine it based on the best load balance and node availability. It is also scalable, currently it processes over 40,000 job submissions a year, many of them parallel. BioHPC integrates distributed cluster resources in a user-transparent way.

FUTURE

The interface source code is freely available. The "packaged" version of the interface can be downloaded from BioHPC.net and installed locally with any Microsoft CCS or HPC 2008 cluster.

We are planning to use our web service interface to integrate BioHPC with the Microsoft Biology Foundation. The pool of applications available through the web service will be expanded and a client will be developed as MS Excel add-in. We currently working on support for 2nd generation sequencing pipelines. The tool will be optimized to take better advantage of multi-core nodes.

ABOUT CBSU

Computational Biology Service Unit (CBSU) of the Cornell University Life Sciences Core Laboratories Center was initiated by the Tri-Institutional collaboration among Cornell University, Well Cornell Medical College, Rockefeller University and Memorial Sloan-Kettering Cancer Center. In February 2006 CBSU became Microsoft HPC Institute charter member. CBSU is Cornell core facility for computational biology. BioHPC development is now partially funded by Microsoft Research and CBSU is a Microsoft Biology Initiative partner.

The BioHPC installation at CBSU is currently using 5 Microsoft Windows based local compute clusters totaling 976 cores. The local nodes use Microsoft Server 2003 with CCS and Microsoft Server 2008 with HPC Server 2008. 80 CPU cores of the remote cluster Athena (located in Redmond, WA) are also available via JSDL, courtesy of Microsoft.