BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 9 2011, pages 1320–1321
doi:10.1093/bioinformatics/btr109

Structural bioinformatics

Advance Access publication March 12, 2011

birgHPC: creating instant computing clusters for bioinformatics and molecular dynamics

Teong Han Chew[1,*], Kwee Hong Joyce-Tan[2], Farizuwana Akma[1] and Mohd Shahir Shamsir[1,*]

[1]Department of Biological Sciences, Faculty of Biosciences and Bioengineering, Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor and [2]School of Biosciences and Biotechnology, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600, UKM Bangi, Selangor, Malaysia

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: birgHPC, a bootable Linux Live CD has been developed to create high-performance clusters for bioinformatics and molecular dynamics studies using any Local Area Network (LAN)-networked computers. birgHPC features automated hardware and slots detection as well as provides a simple job submission interface. The latest versions of GROMACS, NAMD, mpiBLAST and ClustalW-MPI can be run in parallel by simply booting the birgHPC CD or flash drive from the head node, which immediately positions the rest of the PCs on the network as computing nodes. Thus, a temporary, affordable, scalable and high-performance computing environment can be built by non-computing-based researchers using low-cost commodity hardware.

Availability: The birgHPC Live CD and relevant user guide are available for free at http://birg1.fbb.utm.my/birghpc.

Contacts: shahir@utm.my; teonghan@gmail.com

Received on October 21, 2010; revised on January 25, 2011; accepted on February 23, 2011

1 INTRODUCTION

The Beowulf-class computing cluster was a breakthrough in cluster computing because it enabled parallel computing to be performed utilizing low-cost commodity hardware and standard software components compared to other supercomputing solutions (Sterling *et al*., 1995). Typical requirements for building a Beowulf-class cluster involve setting up the number of desired personal computers (PCs) and interconnecting them on the same Local Area Network (LAN), preferably using gigabit ethernet. A similar operating system (OS) and software are then installed on each PC (node) and additional setup is performed for the head node, which is one of the PCs designated to control the computing cluster by taking job submissions and monitoring tasks. While the hardware setup required to build a Beowulf-class computing cluster is rather straightforward, the software setup is challenging to inexperienced researchers and engineers. A Unix-like OS is often chosen as the main OS because they make the implementation of software, such as message passing interface (MPI) libraries, easier. In addition to the OS installation, the configuration of parallel-specific software is also difficult for beginners and non-computing-based researchers, especially those involving the three famous steps: configure, make and make install.

OS and software installations are simplified by using the Linux Live CD method. A Live CD, or its image, can be booted on any PC without affecting the existing users' previously installed OS or data. Several Live CD distributions are currently available with Knoppix (Knopper, 2000) arguably being the most well known. Although early versions of Live CDs were intended for use on a single PC, recent distributions, such as PelicanHPC (successor to ParallelKnoppix) (Creel, 2008) and ClusterKnoppix (clusterknoppix.sw.be) introduced parallel computing capability. Encouraged by the Live CD's simplicity, a number of bioinformatics-based Linux distributions were developed, including Vigyaan (www.vigyaancd.org), Bioslax (www.bioslax.com), Bio-Linux, Quantian (Eddelbuettel, 2003), Open Discovery and DNALinux (Field *et al*., 2006; Rana and Foscarini, 2009). The bioinformatics software included in these distributions are GROMACS (version 3.2.1 in Vigyaan, version 3.2.2 in Quantian and version 3.3 in Open Discovery v2) (Lindahl *et al*., 2006), NCBI's BLAST, ClustalW and others. Only a few bioinformatics-based programs are provided in a Live CD format and are parallel-computing ready, namely Quantian, OpenDiscovery v3 and Knoppix for InterProScan 4.1 High-Throughput Edition (Konishi *et al*., 2006). Open Discovery v2 supports parallel computing but only in the context of a single machine with a multicore processor. However, Quantian supports parallel computing on multiple machines by means of an obsolete OpenMosix kernel. While Open Discovery v3 now provides parallel computing capability, it is not freely available.

We present birgHPC, a Live CD distribution with updated versions of GROMACS, NAMD (Phillips *et al*., 2005), mpiBLAST (Darling *et al*., 2003) and ClustalW-MPI (Li, 2003) that implements a parallel computing capability using OpenMPI and MPICH2, a feature missing from most bioinformatics-based Live CDs.

Our release, birgHPC, is an improved version of PelicanHPC, a live CD that provides Octave parallel computing capability (Creel, 2008). PelicanHPC implements only OpenMPI and requires users to manually define the number of machine slots to utilize the nodes fully, which is time consuming because it requires the user to work with Media Access Control (MAC) addresses in a heterogeneous hardware environment. birgHPC accounts for these shortcomings by including a job submission interface, automatic slot detection and a MPICH2 parallel environment capability.

*To whom correspondence should be addressed.

2 METHODS

The scripts used to build PelicanHPC were modified to build birgHPC. It was augmented with automatic slot detection capabilities and with the ability to make modifications to the host name in order to accommodate the MPICH2 environment. We also included installations of PyMol and Grace from the Debian repository, in addition to allowing manual installations of the latest VMD, fftw libraries, GROMACS, NAMD, mpiBLAST, ClustalW-MPI, OpenMPI and MPICH2 versions from their respective web sites. Some of the original PelicanHPC scripts were incorporated unchanged into birgHPC, such as the cluster configuration script and the netboot image generation script.

3 DISCUSSION

The main feature of birgHPC is its ability to convert PCs on the same LAN into a high-performance computing cluster. PCs in computer labs are often found idle after office hours and during holidays. These untapped resources can be converted within minutes into a computing cluster by booting birgHPC and following the general cluster creation work flow (i.e. setup the head node then the computing nodes). For a PC to act as the head node, it must be able to boot from a CD (birgHPC image) and have a monitor, a mouse and a keyboard. To act as computing nodes, the other PCs need to be bootable from the network. The birghpc_setup script is executed upon the first cluster configuration. When additional nodes are added or removed from the existing cluster, the birg_restart_hpc script is executed. Node statuses can be monitored using the included Ganglia Monitoring System. After the final configuration, users are able to run jobs by placing their files in the /home/user directory, which is shared across the nodes using the network file system (NFS). birgHPC includes a job submission script (/usr/bin/job.sh), which provides users with an interface to run GROMACS, NAMD, mpiBLAST and ClustalW-MPI jobs.

The amount of RAM available is a concern because everything is loaded into RAM. While all the files in the /home/user/ folder are erased from RAM when the head node is rebooted (as is common with many Live CDs), a solution has been implemented in birgHPC which automatically detects and mounts an external storage device as /home (e.g. a flash drive formatted in ext3 with a BIRG label). Testing of birgHPC has been conducted with standard benchmark files and its performance is on par with the normal hard disk-installed computing clusters (Table 1).

Table 1. A performance comparison between a permanent cluster (Ubuntu) and a birgHPC temporary cluster having identical hardware configurations based on the GROMACS with DPPC and Villin benchmarks

Number of processes	DPPC (ns/day)		Villin (ns/day)	
	Permanent cluster	birgHPC temporary cluster	Permanent cluster	birgHPC temporary cluster
2	0.727	0.723	29.799	28.806
4	1.410	1.410	48.010	50.834
6	2.053	2.043	61.727	57.612
8	2.684	2.692	66.745	72.014
10	3.311	3.311	66.475	72.014
12	3.964	3.964	72.014	78.561

4 CONCLUSION

The Linux Live CD, birgHPC, is an improvement over PelicanHPC and Debian Live because birgHPC incorporates capabilities for automatic slot detection, the use of a MPICH2 environment and a job submission interface. The ability of birgHPC to convert LAN-connected PCs into a high-performance computing cluster allows researchers to instantly utilize otherwise idle PCs in laboratories to conduct bioinformatics and molecular dynamics studies with minimal effort. birgHPC is the only freely available, parallel computing-enabled Live CD with the latest versions of GROMACS and NAMD for molecular dynamics simulations.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Universiti Teknologi Malaysia for the facilities provided. Special thanks to Michael Creel for the feedback through PelicanHPC forum.

Funding: Malaysian Government (MOHE FRGS 78531).

Conflict of Interest: none declared.

REFERENCES

Creel,M. (2008) PelicanHPC tutorial. *UFAE and IAE Working Papers*. Unitat de Fonaments de l'Anàlisi Econòmica (UAB) and Institut d'Anàlisi Econòmica (CSIC).

Darling,A.E. *et al*. (2003) The design, implementation, and evaluation of mpiBLAST. In *Proceedings of ClusterWorld*. Citeseer.

Eddelbuettel,D. (2003) Quantian: a scientific computing environment. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*. ISSN 1609-395x.

Field,D. *et al*. (2006) Open software for biologists: from famine to feast. *Nat. Biotechnol.*, **24**, 801–804.

Knopper,K. (2000) Building a self-contained auto-configuring Linux system on an iso9660 filesystem. In *Proceedings of the 4th Annual Linux Showcase and Conference, Atlanta, GA, USA*. The USENIX Association.

Konishi,F. *et al*. (2006) Improving the research environment of high performance computing for non-cluster experts based on knoppix instant computing technology. In *Euro-Par 2006 Parallel Processing*, Springer, pp. 1169–1178.

Li,K.B. (2003) ClustalW-MPI: ClustalW analysis using distributed and parallel computing. *Bioinformatics*, **19**, 1585.

Lindahl,E. *et al*. (2006) GROMACS 3.0: a package for molecular simulation and trajectory analysis. *J. Mol. Model.*, **7**, 306–317.

Phillips,J.C. *et al*. (2005) Scalable molecular dynamics with NAMD. *J. Comput. Chem.*, **26**, 1781–1802.

Rana,A. and Foscarini,F. (2009) Linux distributions for bioinformatics: an update. *EMBnet.news*, **15**, 35–41.

Sterling,T. *et al*. (1995) BEOWULF: a parallel workstation for scientific computation. In *Proceedings of the 24th International Conference on Parallel Processing*. IEEE Computer Society.