BLAST on the Cloud

Christiam Camacho, Chris Cope, Matt Lesko, Thomas Madden
National Center for Biotechnology Information, NIH, Bethesda, MD 20894
{camacho,cope,leskomw,madden}@ncbi.nlm.nih.gov

Introduction

The National Center for Biotechnology Information (NCBI) offers free access to the Basic Local Alignment Search Tool (BLAST) via the World Wide Web [1][2].

In order to provide a fair and consistent level of service, the NCBI limits the size and number of BLAST searches each user can perform. However, as the cost of sequencing decreases [3], users needing to perform large numbers of BLAST searches may find themselves subject to usage limits. Additionally, users with proprietary or custom data sets are currently forced to use the command line BLAST+ [4] interface.

BLAST @ Amazon Web Services

With the aforementioned users' challenges in mind we developed the BLAST Amazon Machine Image (AMI) hosted at Amazon Web Services (AWS). This AMI provides users with the ability to instantiate Linux server(s) at AWS which come pre-configured with:

- the latest release of BLAST+,
- support for a subset of the NCBI BLAST URL API
- a simplified BLAST web page, and
- A File system in User space (FUSE) client that automatically downloads the most recent copy of popular of BLAST databases from NCBI.

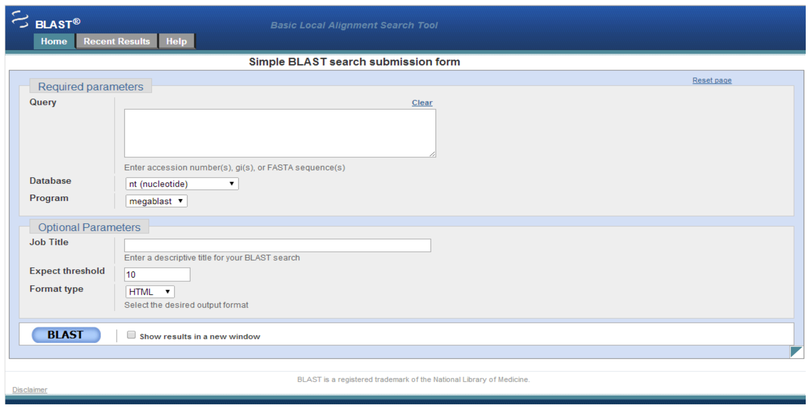

Figure 1: Simplified BLAST web page

System Description

The BLAST AMI is a basic building block to run BLAST on the cloud. It can be used as provided to serve single or multiple users (see Fig. 2) or it can be reused as a modular component in a more elaborate system.

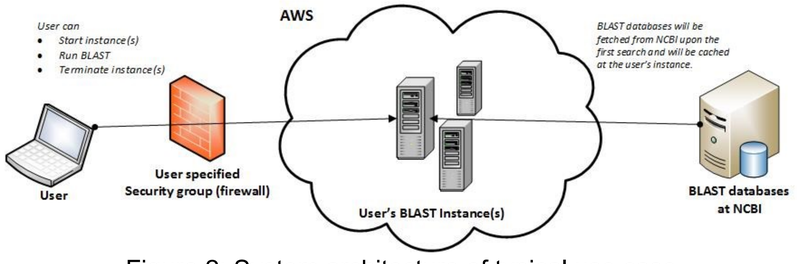

Figure 2: System architecture of typical use case

The BLAST AMI acts both as a front– and back-end server. It accepts BLAST searches via its web interface or via a subset of the NCBI BLAST URL API, facilitating integration into existing workflows which use the public NCBI BLAST service.

BLAST searches are executed immediately after submission in the local host, fetching BLAST databases from NCBI as needed and caching them locally. The users of the AMI do not pay for this network traffic.

Additionally, users can log into their instance(s) and use the BLAST+ command line applications or install additional BLAST databases.

When using the web interface or URL API, the BLAST results are stored in the host's ephemeral storage. These can be viewed, downloaded and deleted via the web interface (Fig. 3); they can be formatted in several popular output formats using the URL API.

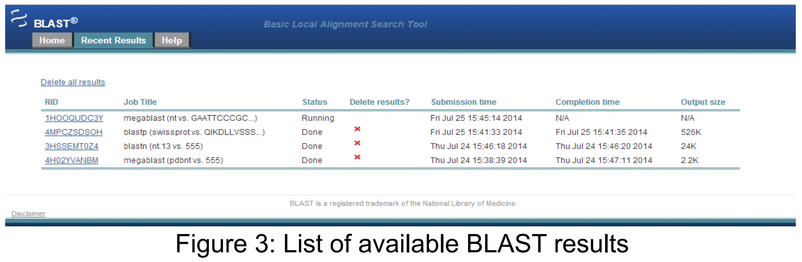

Figure 3: List of available BLAST results

System Performance

Because the BLAST database files are fetched over the network the first time they are accessed, the first BLAST database search has much poorer performance on AWS than on a system having the BLAST databases available locally; however once the databases are locally cached, the performance is comparable (see Fig. 4)

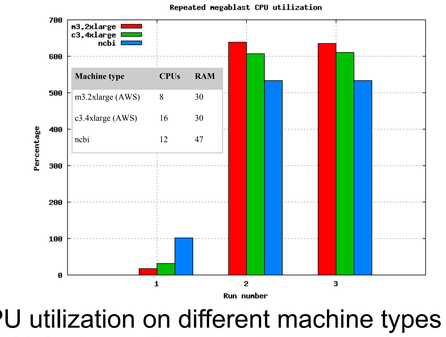

Figure 4: CPU utilization on different machine types when running megablast multiple times. The search was conducted with 11 query sequences totaling 185,780 bases against the nt database from February 2014 using 8 threads and BLAST archive output format. The entry labelled 'ncbi' corresponds to a linux server having the BLAST databases available via NFS, whereas the other two entries corresnpod to machines in the AWS us-east-1a region.

Future work

- Explore distribution and partitioning of BLAST searches among multiple nodes
- Provide a cluster-in-a-box solution to run BLAST

NCBI welcomes feedback and ideas on how to improve on this experiment.
For more information, visit the NCBI BLAST help page using the following QR code:

Acknowledgements

The authors thank Alexey Iskhakov for the FUSE client, Irena Zaretskaya for web development, Eugene Yaschenko, Don Preuss, and Jim Ostell for helpful discussions and guidance.

References

[1] http://blast.ncbi.nlm.nih.gov
[2] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997 Sep 1;25(3389-402)
[3] http://www.genome.gov/sequencingcosts/
[4] Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. BMC Bioinformatics. 2009 Dec 15;10:421.

This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.