Parallel BLAST over Windows HPC Server 2008

Introduction:

BLAST is **B**asic **L**ocal **A**lignment **S**earch **T**ool. Given a biological sequence (DNA/Protein), and a biological database (DNA/Protein), BLAST is used to find regions of similiarity between the submitted query sequence and sequences in the database.

We have implemented WinParallelBLAST program consisting of two modules:

1. **Database segmentation:** This module segments the database according to the number of nodes and format it on each cluster node using the Formatdb BLAST module.
2. **Query distributor:** This module distributes the queries submitted to on the cluster nodes.

Distribution:

1. **Pre-compiled binaries for dividing and formatting database and running BLAST queries.**
2. **Source code:** For the wrapper modules (CLUSTER_Formatdb and CLUSTER_BLASTALL) adapting BLAST for the Windows cluster.

Prerequisites:

1. **Windows HPC Server 2008 RTM version.**
2. **MS-MPI,** Installed directly with HPC component
3. **Visual Studio 2008** (for compiling our source code, if needed)
4. **Executable BLAST file** (www.ncbi.nlm.nih.gov/BLAST/)

Installing BLAST

Obtain BLAST from its web site given above. The installation step of BLAST is so easy just double click on the BLAST icon and it will be installed in the directory which contains the installation source file. In our case we moved the installed packages to D: directory.

The BLAST now is installed in `D:\Blastout the excution files of BLAST and FORMATDB is located in D:\Blastout\bin`

Note: `Our programs for parallel BLAST` **CLUSTER_Formatdb.c** and **CLUSTER_BLASTALL.c** assume that **BLAST** and **formatdb** are in `D:\Blastout\bin .` You can change the paths from the code.

Running the Parallel BLAST program on the cluster

Formatting the database

First you need to go to **CLUSTER_formatdb.exe** and create a shared directory to be accessible through all cluster nodes.

Steps:

1. Move the **CLUSTER_formatdb.exe** to the shared folder
2. Then create input file, containing input files to the job scheduler, because **CLUSTER_formatdb** need two inputs
   a. Type of the database in case of protein the choice will be 1 and in case of DNA the choice will be 2.
   b. The path of the database.

The input file is a text file with two lines: the first one will be the Database type and the second one will be the path of the database. Here is an example of an input file, where the first line specifies that it is DNA file and the second line points to the database.

```
1
\workingdir\shareddir\MyDatabase
```

3. Run this command from the Command prompt

```
>job submit /numofnodes:4 /workdir:\\H-Node\Users\Hishama\Desktop\Run\
/stdin:input.txt /stdout:_out.txt /stderr:error.txt mpiexec –
machinefile hosts.txt –n 4 cLUSTER_formatdb.exe
```

This command will run the **CLUSTER_formatdb.exe** on all nodes listed in the **hosts.txt** file. The hosts file contains the name of nodes in my cluster in our case it contains **H-Node,C-Node1,C-Node2,C-Node3** but each one is in separate line.

As mentioned before, **CLUSTER_formatdb.exe** divide the database on each node and make the call the **formatdb** of BLAST to format the sub-database on each node. If the job is submitted correctly, then you will find on each node number of files in **C:\BLASTDB** these files are the part of database after being formatted on this node .The same will be found on other nodes.

Now you have divided the database on all nodes of the cluster, and you are ready to submit queries to these portions of the database

T o submit one query

```
>job submit /numofnodes:4 /workdir:\\H-Node\Users\Hishama\Desktop\Run\
/stdin:input.txt /stdout:_out.txt /stderr:_err.txt mpiexec –
machinefile hosts.txt –n 4 BLAST_ALL.exe
```

In this time the input file will contain three things

a. The path of the original database
b. The path of the query file
c. Type of search
    1. if DataBase is Protein and Query is Protien
    2. if DataBase is Nucleotide and Query is Nucleotide
    3. if DataBase is Protein and Query is Nucleotide
    4. if DataBase is Nucleotide and Query is Protein

Each input has to be in separated line. Here is an example of an input file, where we have the database in the file Mydatabase.

```
C:\Mydatabase
C:\query.txt
1
```

The Results of the BLAST will be output in **`C:\Blast-result\`**

Compiling our wrapper modules

In order to compile the two files you will need Visual studio 2008 and the MS-MPI library.

1. Open visual studio 2008 -> File -> New ->Project->Visual C++-> Win32 console application->and enter the name of the project **BLAST_Formatdb**->**ok**

2. New windows will appear click **Next->**choose console application and empty project then click **Finish.**

3. Right Click on the project from **Solution Explorer** then **Add-> Existing Item** and choose the **Formatdb.c** file.

4. Now We need to include the MPI library. Right click on the project select **properties** then **C\C++** then in **Additional Include Directory** enter the path of **msmpi.h** which is mainly located in **C:\Program Files\Microsoft HPC Pack 2008 SDK\ include** as shown in figure(1).

5. Choose **Advanced** then **compile as** and choose **C Code** as shown in figure(2).

6**.** Choose linker then in **Additional library Directory** enter the path of **msmpi.lib** which is mainly in **C:\Program Files\Microsoft HPC Pack 2008 SDK\lib** as shown in figure(3).

7. Choose **Input** then in **Additional Dependencies** write **msmpi.lib** as shown in figure (4)**.**

8. For large data files you will need to increase the memory choose **System** then modify the value **in Heap Reverse Size, Heap commit Size, Stack Reservesize, Stack commit size according to your need** as shown in figure(5) **.**

9. Now you can compile and run the program.

10. Same steps will be done for the **BLAST.c** file

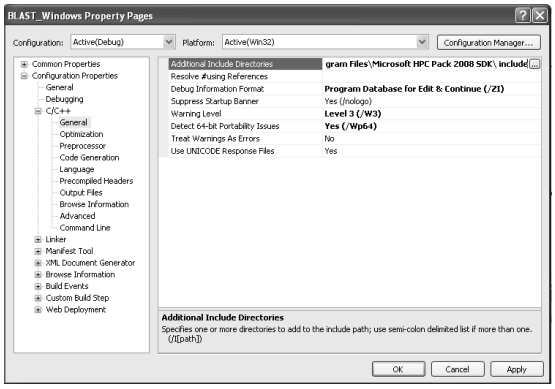

Figure 1 Compile

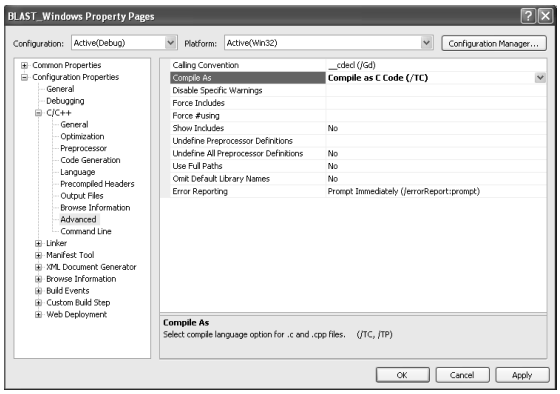

Figure 2 Compile

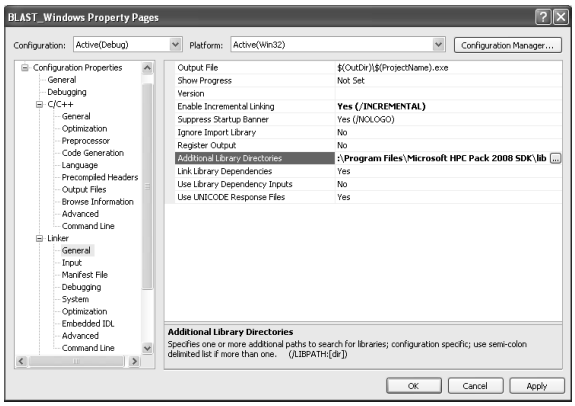

Figure 3 Compile

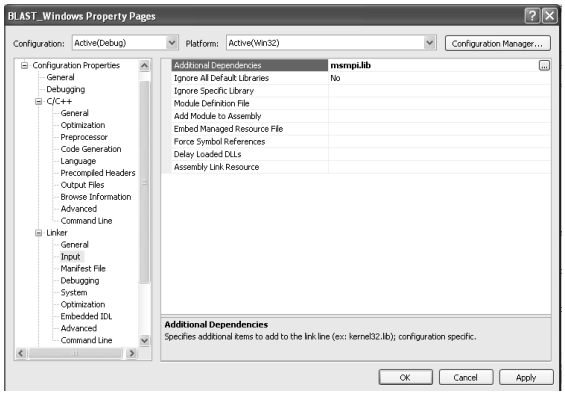

Figure 4 Compile

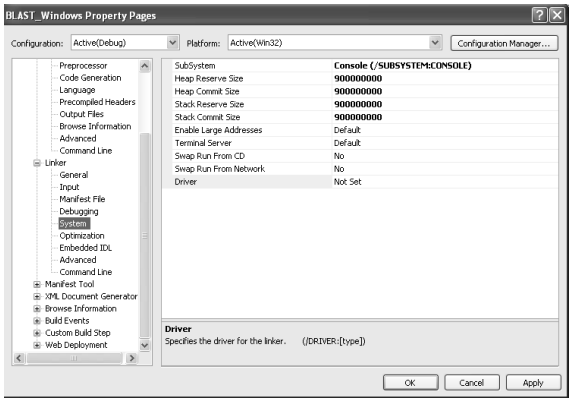

Figure 5 Stack Allocation

The executable file will be in the Debug folder.