Introduction
The goal of this project is to create a High Performance Computing (HPC) infrastructure for large-scale genome and post-genome era data analysis at SIUC. The parallelized computational pipeline includes multiple genome assemblers, user web interface, SQL database and statistical and graphical output to support fast and convenient genome analysis for biologists. Various computing and memory intensive tasks can be integrated and executed in an automated way to reduce human interference.

Objectives

**Obj 1.** To construct the parallelized process on a distributed system which significantly speeds up the assembly and analysis of the genome.

**Obj 2.** To provide the most accurate results possible. Running the data in parallel through many different assemblers will likely produce very different outputs.

**Obj 3.** To create a web interface for easy integration with the parallelized system.

Architectural Design

From input to output

BigDog

BigDog is SIU high performance computing (HPC) and based on Cisco Unified Computing System (UCS) solution and includes 40 C-Series nodes and a Nexus 9000 series switch with 70TB of storage on the front-end machine, and two GPU nodes with dual NVIDIA Tesla K40 cards. BigDog has forty compute nodes with either 64GB or 768GB of RAM and twenty cores each, giving the entire cluster a theoretical max of 34.7 TeraFLOPS. BigDog is used by researchers in material science, computer science, biology and medicine at SIU.

BigDog Configuration

<table>
<thead>
<tr>
<th>Machine</th>
<th>Node</th>
<th>Initial XEON E5 Processor Per Node</th>
<th>Core Per Node</th>
<th>RAM Per Node</th>
<th>TBO Per Node</th>
<th>NVIDIA K40 GPU</th>
<th>CUDA Core Per Node</th>
<th>GDDR5 Memory Per Node</th>
<th>Ports and Connectors</th>
</tr>
</thead>
</table>
| Nexus 6300 1 | 40 port 10G & 6 port 1G | 264 | 64GB | 1T
| UCS C-220 16 | 20 | 64GB | 1T
| UCS C-220 2 | 20 | 768GB | 4T
| UCS C-240 2 | 20 | 64GB | 1T | 2 | 2860 | 12GB | 48 port 10G & 6 port 40 |

Testing Results

- Initial tests on moderately sized data showed a decrease in time from 3 days on a typical computer to 3 hours on the high memory nodes of the BigDog.
- We tested two desktops and BigDog with three sets of data and had three runs of the data. Desktop 1 could not handle the tasks in the Run 2 and Run 3. Although Desktop 2 could do the tasks, BigDog would make the run more efficiently.

Conclusions

- The project increased efficiency, accuracy, and accessibility for genomic research.
- It will make it easier for researchers in genome assembly and analysis at SIUC.

Acknowledgements

- Departments of Computer Science, Information Technology and Plant Biology.

Reference