Title of workshop:

Building Linux Clusters for Bioinformatics Applications.

Teacher:

Mr. Jacek Nowak.
Team Leader of Bioinformatics Group, Plant Biotechnology Institute, National Research Council Canada. Designed, implemented and maintained a 64 processor, Beowulf class supercomputer (cluster). The cluster provides computing resource in support of high throughput data acquisition and analysis arising from Agricultural Genomics program at NRC.

Tutorial presentation:

Effective analysis of the rapidly expanding genome science data sets necessitate that bioinformatics efforts be supported by: cost effective and high performance computer platforms and solutions, reliable and effective storage and analysis mechanisms, new data analysis tools.
The goal of the proposed workshop is to provide understanding of existing clustering technologies, methodologies and basic setup principles, as applied to bioinformatics research. It is intended as a mini “how-to” on the basic design, implementation, maintenance and software support of Linux based, Beowulf class supercomputer.
The workshop will consist of a series of slides presentations, supported by detailed information in form of handouts. The contents will include publicly available information as well as personal experiences in building Linux clusters for bioinformatics applications.
The workshop will follow an outline of recommended companion book:
D.H.M Spector, Building Linux Clusters, O’Reilly, 2000.

Intended audience:

This workshop is intended for those interested in designing, building, maintaining and programming a Beowulf class computer cluster in support of bioinformatics efforts in respective research organizations including IT administrators, researchers as well as students.
The workshop will be taught at the introductory and intermediate level, and requires basic understanding of computer technologies, networking and programming as well as understanding of bioinformatics sciences principles including sequence comparisons and sequence similarity database searches (e.g. BLAST).

Detailed outline:

1. Introductory information
An overview of recent developments in life science research, particularly in high-throughput gene sequencing and function analysis, and the impact on the bioinformatics. Comparison of existing hardware/software solutions as applied to genomic studies. A brief history of distributed computing.

2. Basic concepts: clustering, networking, parallel programming
This section will cover basic concepts involved in parallel computing, multiprocessing, cluster solutions and programming including hardware and software parallelism.
Introduction to network and networking terminology including protocols, interfaces, medium and bandwidth, and exploration of IP addressing technologies and routing topologies available for parallel computing environments.

3. Designing clusters
Exploration of hardware choices available, discussion on hardware requirements for different platforms and hardware architectures, interaction and limitations of heterogeneous cluster solutions, discussion on the proper choice of hardware and platforms for different types of computing problems. This section will discuss characteristics of different memory, processor and I/O types on cluster configurations and styles.
Exploration of networking architectures, designs and implementations with the emphasis on network performance in a parallel environment. Discussion on network interconnect selection and recommended configurations.
Scalability, data access and security considerations will also be discussed.

4. Building clusters
Essential information on selection of location, space, power, access and cooling methods and requirements.
Practical notes on requirements and assembly of custom cluster components including server (master node), computing nodes, network setup, wiring and shelving including overview of available designs and choice of systemboard, memory, processor, disk, networking and enclosure components.
Discussion on proper selection of computer parts vendors, including delivery and warranty issues. Discussion of pros and cons of contracted out vs. self assembly approaches to building and selection of the cluster.

5. Cluster software installation and configuration
Overview of cluster software installation including planning and layout of the system setup. This section will include design decisions vs. computational needs, distribution choices, booting methods, hard disk setup and local storage methods. Detailed information on the software and setup requirements of different styles of computing nodes and server systems, as well as network

configuration and individual packages and applications installation will be discussed.

6. Managing clusters

This section will discuss and present a variety of tools, software, methods and concepts required for the management of critical cluster components and basic maintenance. Exploration of common solutions and custom procedures dealing with the management of nodes, users, projects, updates and security. Discussion of resource allocation and assignment, job queuing, quote management and process accounting as well as software and operating system upgrades and maintenance.

7. Bioinformatics applications and cluster programming

Overview of existing bioinformatics applications including discussion of licensing issues and availability for use in genomic studies, with a particular emphasis on parallel/distributed versions. Exploration of methods and approaches aiming at parallelisation of existing applications.

This section will discuss tools and libraries for working in cluster environment including distributed file systems, messaging (PVM, MPI) and queuing tools and approaches as well as system extensions(BPROC, MOSIX).

8. Optimization of cluster design for bioinformatics research

Much of the sequence analysis involves comparing a query sequence or a pattern to a reference database. In general, the time required to complete such a search is directly proportional to the size of the database. One method of the database search algorithm optimization, as applied to the distributed cluster environment and leading to the overall improvement of the performance will be discussed.

9. Links to available resources

A list of the most common and useful links to the sources of information on the topics of clustered computing, programming, networking and hardware design, and parallel bioinformatics applications including books, magazines and periodicals, conferences and online resources will be presented and discussed.