PM14: ISMB2006 Tutorial

Title: Maximize Genomics Throughput with Data-Activated Processing

Topic Area:
- Database and Data Integration (25%)
- Other: Data Streaming for Genomics (75%)

Main Presenter:

- Dr. Stephen C. Perrenod
- eXludus Technologies Inc.
- 19925 Stevens Creek Boulevard, Cupertino, CA 94015
- Stephen.perrenod@exludus.com
- W: +1-408-725-7571 Cell: +1-650-867-1532
- +1-253-736-0155
- www.exludus.com

- Teaching experience: 25 years of presenting HPC technology to audiences Small and large (up to 500). Harvard teaching assistant (Astronomy)
- Lepton Photon 2003 "Data Grid" evening, Fermilab, Illinois http://conferences.fnal.gov/lp2003/bulletins/grid_new.html

Second presenter:

- Dr. Ulrich Meier
- Sun Microsystems GmbH
- Brandenburger Str.2, 40880 Ratingen, Germany
- Ulrich.Meier@sun.com
- +49 2302 780185
- +49 2302 780186
- www.sun.com/lifesciences

- Teaching experience: 20 years of presenting HPC and Graphics technologies to scientific audiences.
- Earlier tutorial presentations: GCB 2005, Sun HPC Consortium.

50-word abstract: Please provide a brief explanatory abstract here for advertising your tutorial.

Effective scaling and maximal throughput for clusters is an ongoing issue for computational bioinformatics workloads. Data volumes are doubling every year. How do we feed all this data into clusters? Data-activated processing allows you to get more work through your cluster system. In this tutorial we will present an approach to workload distribution which is data-centric, rather than process-centric.

Tutorial level: Introductory

Prior knowledge required: Basic knowledge of genomics processing workflows and familiarity with cluster and grid computing.

Suitability of this tutorial for ISMB:
Increasingly, bioinformatics workloads are handled on clusters, and throughput is often severely degraded due to data bottlenecks in file severs and networks supporting the clusters. The approach covered in this tutorial is designed to eliminate data bottlenecks found in bioinformatics workflows.

- Timeliness – this is a new approach to addressing the severe data management issues being faced
- Audience – Researchers and bioinformatics system administrators
- Cutting-edge technology – this is brand-new technology which can be applied across a range of application workloads
- The methodology is suitable to be applied to a large number of bioinformatics problems.

Profile of Dr. Perrenod

- My interests are in maximizing return on investment from high performance computing architectures. I have 25 years experience working with HPC environments.
- Teaching: Harvard T.A. in astronomy; 25 years of presentations and lectures on HPC and Grid technology to large and small audiences. Workshop at Lepton Photon 2003, August on Data Grid. Workshop on Data Grids and Building the Digital Campus at Canadian Bioinformatics Resource, Halifax, May 2004.

Profile of Dr. Meier

- Life Sciences application tuning and parallelization. Grand Challenge problems in bioinformatics and computational chemistry
- Prior teaching, assistance lecturer at the University of Bochum, various workshops on performance tuning, compiler usage, parallelization for shared and distributed memory and grid computing.
- 20+ years of subject matter industry expertise.

Tutorial Outline:

- The outline should be a table of contents of the tutorial, with a few keywords for each section, and with a rough estimate of the time spent on each.

1. The data crisis (0.3 hours)
    a. Data volumes doubling, bottlenecks and queuing issues
    b. Genomics example (BLAST)

2. Today the world is dominated by process-centric approaches.
    a. What is process-centric workflow, why do we use it? (0.3)
    b. What problems does this approach raise? (0.3)

3. What is data-activated processing? (0.3)

4. How can data-activated processing be implemented? (0.6)

   a. Broadcasting to all or many nodes
   b. Data sharing

5. What results can be achieved? (0.4)
   a. Broadcasting performance
   b. NCBI-BLAST results (human, rice genomes)

6. How do we integrate with existing environments? (0.3)

   a. With existing file systems
   b. With existing workload management systems

7. N1 Grid Engine features for high throughput (0.5)

8. Discussion (0.5)