Microsoft Windows Compute Cluster Server 2003

The Power of Distributed Computing in Cancer Research

Supercomputing 2005 Keynote Demonstration

Overview

Current methods for diagnosing cancer often require expensive and invasive biopsies that don't always facilitate early detection, are dangerous for the patient, and are time and resource intensive for hospitals and doctors. A promising alternative is a blood test that could identify the early stages of cancer—greatly improving the prognosis for many patients. One hurdle has been that identifying the differences between normal, healthy, blood samples and blood samples from patients with cancer requires analyzing and correlating results from hundreds or thousands of blood tests, each of which contain hundreds of thousands of different molecules. Until now, this type of analysis could only be performed on supercomputers costing millions of dollars.

Today it is possible to create high-performance computing (HPC) clusters using cost-effective x64-based servers. These HPC clusters have the computational power of a supercomputer at a fraction of the cost. Laboratories of all sizes can now solve real-world problems that were beyond their resources only a few years ago.

This demonstration shows how arrays of standard x64-based servers running Microsoft® Windows® Compute Cluster Server 2003 (CCS) can provide the parallel computational capabilities to support the research and diagnostic needs of the medical and scientific community. Because CCS is easy to deploy, use, and manage, and supports cluster sizes ranging from the personal cluster to the very large cluster, it reduces time-to-insight, boosts end-user productivity, and allows scientists to spend less time setting up IT and more on research.

In this demonstration, Microsoft® SQL Server™2005 Integration Services are used to drive online data acquisition from the FDA-NCI Clinical Proteomics Program Web site. The SQL Server™ 2005 Integration Services automated workflow converts the data from a flat file into a structured form and then preprocesses the data using a CCS cluster. With its seamless integration with the MATLAB® technical computing platform and its user-friendly interface, CCS makes compute clusters easier to use and reduces time-to-insight for research scientists. Finally, the full data set is processed using a remote large scale CCS cluster and a Linux™ cluster through Platform™ LSF® powered by Platform Enterprise Grid Orchestrator™ (EGO), as shown in Figure 1.

The Core Analysis Process

Early detection is a major factor in successful treatment of most types of cancer. However, current diagnostic methods generally require invasive and expensive biopsies. A blood test that could identify early stages of cancer could greatly improve outcomes for many patients. Proteomics—the study of all the proteins in a cell, tissue, or organism—is a very promising approach to "fingerprinting" cancer cells with a blood test. Recent improvements in protein mass spectrometry make proteomic fingerprinting ideal for identifying the biomarkers of disease.

The publicly available data sets from the National Cancer Institute include several hundred mass spectra of blood serum samples from cancer patients and similar numbers of spectra from healthy control patients. Each spectra consists of several hundred thousand data points, each of which corresponds to the relative amount of a particular protein in the sample. Comparative analysis of protein mass spectrometry data is a highly parallel task that lends itself particularly well to distribution across a compute cluster. The analysis of these large data sets takes advantage of the data workflow capabilities in SQL Server™ 2005 and the resource allocation capabilities of CCS. Initially, the analysis utilizes a personal compute cluster for interactive prototyping of algorithms for feature selection with MATLAB. It then processes the full data sets using a large, cross-platform, distributed compute cluster to identify the optimum set of features to be used for classification. The data points in the optimal feature set are then used to look up proteins for further study as potential biomarkers.

> "Proteomics data are being collected at a faster pace than the ability of the researchers to validate, interpret, and integrate them with other known data. There is a great need to make data portable and comparable. Software tools are needed in all areas of data analysis, including data collection, storage, searching, analysis, classification, management, archiving, and retrieval."
>
> —National Cancer Institute

Computational Analysis

The tasks are as follows:

1. Create an automated workflow in SQL Server™ 2005 to download the data sets from the Web, store the structured data in a database, and execute the MATLAB-prototyped pre-processing scripts (background correction, normalization, and correction for miscalibration).
2. Load the data from SQL Server™ directly into MATLAB. Using the familiar MATLAB programming language, interactively tune the selection algorithm to identify a subset of features

The Power of Distributed Computing in Cancer Research

Microsoft Windows Compute Cluster Server 2003 Demonstration

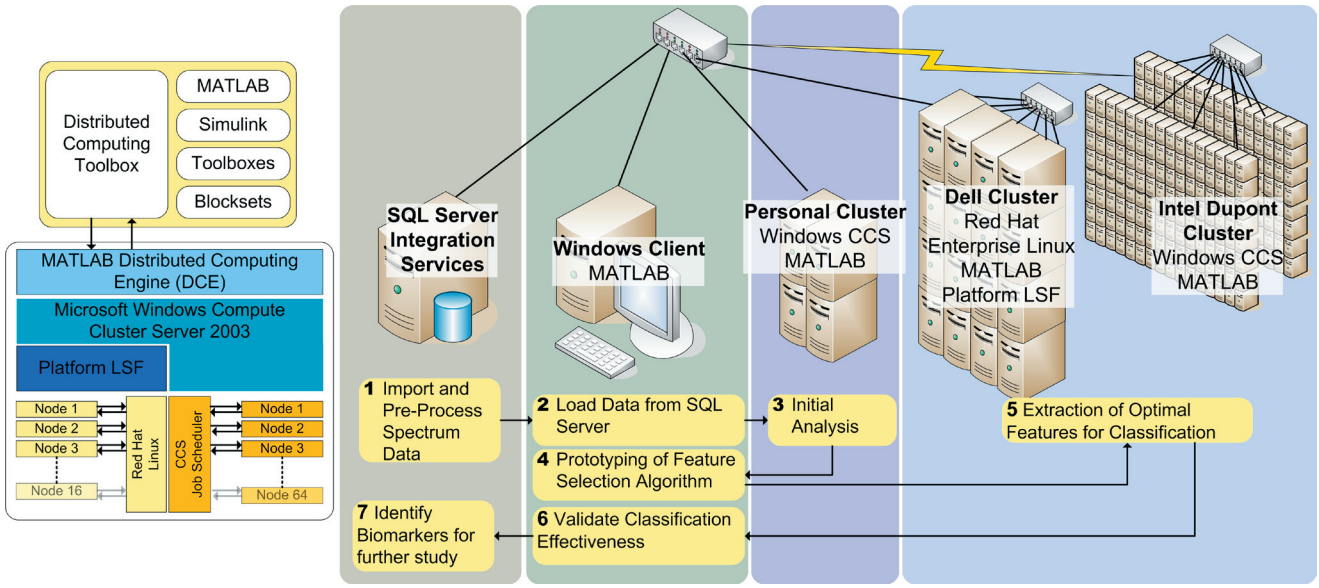

Figure 1. The demonstration shows the power of CCS in distributed computing

that can classify the new spectra. The M-files are seamlessly submitted and run from within MATLAB on a personal cluster running CCS, allowing interactive prototyping of the algorithms to be used for feature selection.

3. After a desired algorithm is chosen, use MATLAB to submit a job that will run the algorithm on the full data set to extract an optimal set of features to be used for classification. CCS, running on the personal cluster, forwards the large job to a remote cluster (also running CCS), and, through Platform LSF, to a Linux cluster.

Computing Infrastructure

The demonstration uses three distinct compute clusters:

- *Personal Compute Cluster*. The personal compute cluster shown during the demonstration is a concept platform developed by Ciara Technologies with assistance from Intel® Corporation. The personal compute cluster consists of four dual-core processors in a deskside system designed to be used in an office environment for solving computationally intensive problems. The system consists of 4 Intel processor-based motherboards, each with a dual-core 64-bit Intel® Xeon® processor and 2 gigabytes (GB) of RAM, all connected by a built-in gigabit Ethernet switch. The cluster is powered by CCS and during the demonstration, it ran MATLAB® MathWorks Release 14 SP3 and Distributed Computing Toolbox V2.

- *Large Remote Compute Cluster*. The HPC cluster, powered by Dual-Core Intel Xeon processors, is located at Intel Corporation's Remote Access facility in DuPont, Washington. It has a total configuration of 128 nodes consisting of dual-core Intel Xeon processors providing a total of 512 cores. The system was reconfigured for use in this demonstration to provide Microsoft with 256 cores for computation. The cluster was configured as a 64-node cluster, with each node consisting of 2 dual-core 64-bit Intel Xeon processors, 8 GB RAM, Gigabit Ethernet and SilverStorm Technologies™ InfiniBand adapters. The Intel dual-core HPC cluster was accessed directly over the SCinet InfiniBand network. The cluster is powered by CCS and runs MATLAB® MathWorks Release 14 SP3 and Distributed Computing Toolbox V2.

- *Linux Compute Cluster*. The Linux 32 CPU cluster is provided by Dell™ Inc. and consists of 16 PowerEdge™ SC1425 1U servers, featuring two 64-bit Intel Xeon processors, 2GB RAM, and gigabit Ethernet connectivity. This cluster is connected to the rest of the demonstration network with a Cisco® Catalyst® 4948 switch. The cluster runs Red Hat® Enterprise Linux® 4, Platform LSF 6.2, and MATLAB MathWorks Release 14 SP3 and Distributed Computing Toolbox V2.

The Results

This demonstration shows the power of distributed clusters of cost-effective industry standard servers to perform the complex data analysis necessary for proteomic fingerprinting. CCS enables scientists and other end-users to improve productivity by allowing seamless access from the workstation to structured data stores, small deskside clusters for interactive analysis, and large heterogeneous pools of computing resources for detailed studies.

Windows Compute Cluster Server 2003 is a powerful solution that is designed to accelerate time-to-insight by providing an HPC platform that is simple to deploy, operate, and integrate with existing infrastructure and tools. Explore CCS at: http://www.microsoft.com/hpc today.

Microsoft Corporation • One Microsoft Way • Redmond, WA 98052-6399 • USA
Version 1.0
0395 Part No. 098-00000

Microsoft®
Your potential. Our passion.™