Seamlessly Scaling a Desktop MATLAB Application to an Experimental TeraGrid Resource

Zoya Dimitrova1, David Camp0, Elizabeth Neuhau2, Yari Khudyakov1, Nate Woody1, David Lifka2

1Division of Viral Hepatitis, Centers for Disease Control and Prevention; 2Scientific Computing Activity, Division of Science Gateway Programs, Centers for Disease Control and Prevention; 2Cornell University Center for Advanced Computing; Corresponding author: lifka@cac.cornell.edu

Introduction

The increasing pervasiveness of distributed and parallel computing in a variety of fields has resulted in the need for a computer environment that will support the needs of the many researchers who do not have access to high performance computing facilities. Distributed and parallel computing is not a new concept, but it has become more prevalent with the increasing power of many-core and multi-core machines. The work presented here is a collaborative effort between researchers at the Centers for Disease Control and Prevention and the Cornell University Center for Advanced Computing (CAC) to develop an efficient and user-friendly system that will enable researchers to seamlessly scale their applications from desktops to remote high performance computing resources.

Abstract

MATLAB® was used to study a number of services needed to enable the client to submit the request to the TeraGrid through the gateway and finally to perform the required computation on the remote system. This enabled a great reduction in HCV simulation times. The MATLAB Distributed Computing Server running on this resource provides parallel and distributed computational services to interactive desktop users and Science Gateways. We have implemented a system that allows a researcher to scale from their desktop to a remote system transparently and with little knowledge of parallel programming. The integration was accomplished by providing a Java-based integration with the MATLAB Parallel Computing Toolbox. The system hosted at the Cornell Center for Advanced Computing (CAC) is composed of a Dell blade cluster running Microsoft Windows HPC Server 2008 and an 8TB network storage device which is duplicated. To this end, not only will the Cornell CAC support TeraGrid customers with this resource, CAC will also be used to test that our approach is indeed user-friendly. We would also like to thank our colleagues at the Centers for Disease Control for their efforts and support that made this resource available. This research is sponsored by the National Science Foundation (Award #0844032), Dell, Intel, Microsoft, and The MathWorks. We would also like to thank our colleagues at the Centers for Disease Control for their efforts and support that made this resource available.

System Architecture

The work presented here is a collaborative effort between researchers at the Centers for Disease Control and Prevention and the Cornell University Center for Advanced Computing (CAC) to develop an efficient and user-friendly system that will enable researchers to seamlessly scale their applications from desktops to remote high performance computing resources.

Results

There are many different sources of variation in amino acid sequences (such as the dataset of HCV genotype 1b). Different sources of variation might introduce a bias in the results, but this can be controlled by the use of appropriate statistical methods. In this paper, we use a parallelized permutation test to assess the statistical significance of the observed differences. We use the JGlobus CoG Platform system which provides a seamless access to computing resources at other institutions through a Grid infrastructure service. The JGlobus CoG Platform system is built on top of the JGlobus Toolkit 4.0, which is a collection of open-source libraries written in C/C++.

Methods Overview

Seamlessly Scaling a Desktop MATLAB Application to an Experimental TeraGrid Resource

Zoya Dimitrova1, David Camp0, Elizabeth Neuhau2, Yari Khudyakov1, Nate Woody1, David Lifka2

1Division of Viral Hepatitis, Centers for Disease Control and Prevention; 2Scientific Computing Activity, Division of Science Gateway Programs, Centers for Disease Control and Prevention; 2Cornell University Center for Advanced Computing; Corresponding author: lifka@cac.cornell.edu

Introduction

The increasing pervasiveness of distributed and parallel computing in a variety of fields has resulted in the need for a computer environment that will support the needs of the many researchers who do not have access to high performance computing facilities. Distributed and parallel computing is not a new concept, but it has become more prevalent with the increasing power of many-core and multi-core machines. The work presented here is a collaborative effort between researchers at the Centers for Disease Control and Prevention and the Cornell University Center for Advanced Computing (CAC) to develop an efficient and user-friendly system that will enable researchers to seamlessly scale their applications from desktops to remote high performance computing resources.

Abstract

MATLAB® was used to study a number of services needed to enable the client to submit the request to the TeraGrid through the gateway and finally to perform the required computation on the remote system. This enabled a great reduction in HCV simulation times. The MATLAB Distributed Computing Server running on this resource provides parallel and distributed computational services to interactive desktop users and Science Gateways. We have implemented a system that allows a researcher to scale from their desktop to a remote system transparently and with little knowledge of parallel programming. The integration was accomplished by providing a Java-based integration with the MATLAB Parallel Computing Toolbox. The system hosted at the Cornell Center for Advanced Computing (CAC) is composed of a Dell blade cluster running Microsoft Windows HPC Server 2008 and an 8TB network storage device which is duplicated. To this end, not only will the Cornell CAC support TeraGrid customers with this resource, CAC will also be used to test that our approach is indeed user-friendly. We would also like to thank our colleagues at the Centers for Disease Control for their efforts and support that made this resource available. This research is sponsored by the National Science Foundation (Award #0844032), Dell, Intel, Microsoft, and The MathWorks. We would also like to thank our colleagues at the Centers for Disease Control for their efforts and support that made this resource available.

System Architecture

The work presented here is a collaborative effort between researchers at the Centers for Disease Control and Prevention and the Cornell University Center for Advanced Computing (CAC) to develop an efficient and user-friendly system that will enable researchers to seamlessly scale their applications from desktops to remote high performance computing resources.

Results

There are many different sources of variation in amino acid sequences (such as the dataset of HCV genotype 1b). Different sources of variation might introduce a bias in the results, but this can be controlled by the use of appropriate statistical methods. In this paper, we use a parallelized permutation test to assess the statistical significance of the observed differences. We use the JGlobus CoG Platform system which provides a seamless access to computing resources at other institutions through a Grid infrastructure service. The JGlobus CoG Platform system is built on top of the JGlobus Toolkit 4.0, which is a collection of open-source libraries written in C/C++.

Methods Overview

Seamlessly Scaling a Desktop MATLAB Application to an Experimental TeraGrid Resource

Zoya Dimitrova1, David Camp0, Elizabeth Neuhau2, Yari Khudyakov1, Nate Woody1, David Lifka2

1Division of Viral Hepatitis, Centers for Disease Control and Prevention; 2Scientific Computing Activity, Division of Science Gateway Programs, Centers for Disease Control and Prevention; 2Cornell University Center for Advanced Computing; Corresponding author: lifka@cac.cornell.edu

Introduction

The increasing pervasiveness of distributed and parallel computing in a variety of fields has resulted in the need for a computer environment that will support the needs of the many researchers who do not have access to high performance computing facilities. Distributed and parallel computing is not a new concept, but it has become more prevalent with the increasing power of many-core and multi-core machines. The work presented here is a collaborative effort between researchers at the Centers for Disease Control and Prevention and the Cornell University Center for Advanced Computing (CAC) to develop an efficient and user-friendly system that will enable researchers to seamlessly scale their applications from desktops to remote high performance computing resources.

Abstract

MATLAB® was used to study a number of services needed to enable the client to submit the request to the TeraGrid through the gateway and finally to perform the required computation on the remote system. This enabled a great reduction in HCV simulation times. The MATLAB Distributed Computing Server running on this resource provides parallel and distributed computational services to interactive desktop users and Science Gateways. We have implemented a system that allows a researcher to scale from their desktop to a remote system transparently and with little knowledge of parallel programming. The integration was accomplished by providing a Java-based integration with the MATLAB Parallel Computing Toolbox. The system hosted at the Cornell Center for Advanced Computing (CAC) is composed of a Dell blade cluster running Microsoft Windows HPC Server 2008 and an 8TB network storage device which is duplicated. To this end, not only will the Cornell CAC support TeraGrid customers with this resource, CAC will also be used to test that our approach is indeed user-friendly. We would also like to thank our colleagues at the Centers for Disease Control for their efforts and support that made this resource available. This research is sponsored by the National Science Foundation (Award #0844032), Dell, Intel, Microsoft, and The MathWorks. We would also like to thank our colleagues at the Centers for Disease Control for their efforts and support that made this resource available.

System Architecture

The work presented here is a collaborative effort between researchers at the Centers for Disease Control and Prevention and the Cornell University Center for Advanced Computing (CAC) to develop an efficient and user-friendly system that will enable researchers to seamlessly scale their applications from desktops to remote high performance computing resources.

Results

There are many different sources of variation in amino acid sequences (such as the dataset of HCV genotype 1b). Different sources of variation might introduce a bias in the results, but this can be controlled by the use of appropriate statistical methods. In this paper, we use a parallelized permutation test to assess the statistical significance of the observed differences. We use the JGlobus CoG Platform system which provides a seamless access to computing resources at other institutions through a Grid infrastructure service. The JGlobus CoG Platform system is built on top of the JGlobus Toolkit 4.0, which is a collection of open-source libraries written in C/C++.

Methods Overview

Seamlessly Scaling a Desktop MATLAB Application to an Experimental TeraGrid Resource

Zoya Dimitrova1, David Camp0, Elizabeth Neuhau2, Yari Khudyakov1, Nate Woody1, David Lifka2

1Division of Viral Hepatitis, Centers for Disease Control and Prevention; 2Scientific Computing Activity, Division of Science Gateway Programs, Centers for Disease Control and Prevention; 2Cornell University Center for Advanced Computing; Corresponding author: lifka@cac.cornell.edu

Introduction

The increasing pervasiveness of distributed and parallel computing in a variety of fields has resulted in the need for a computer environment that will support the needs of the many researchers who do not have access to high performance computing facilities. Distributed and parallel computing is not a new concept, but it has become more prevalent with the increasing power of many-core and multi-core machines. The work presented here is a collaborative effort between researchers at the Centers for Disease Control and Prevention and the Cornell University Center for Advanced Computing (CAC) to develop an efficient and user-friendly system that will enable researchers to seamlessly scale their applications from desktops to remote high performance computing resources.

Abstract

MATLAB® was used to study a number of services needed to enable the client to submit the request to the TeraGrid through the gateway and finally to perform the required computation on the remote system. This enabled a great reduction in HCV simulation times. The MATLAB Distributed Computing Server running on this resource provides parallel and distributed computational services to interactive desktop users and Science Gateways. We have implemented a system that allows a researcher to scale from their desktop to a remote system transparently and with little knowledge of parallel programming. The integration was accomplished by providing a Java-based integration with the MATLAB Parallel Computing Toolbox. The system hosted at the Cornell Center for Advanced Computing (CAC) is composed of a Dell blade cluster running Microsoft Windows HPC Server 2008 and an 8TB network storage device which is duplicated. To this end, not only will the Cornell CAC support TeraGrid customers with this resource, CAC will also be used to test that our approach is indeed user-friendly. We would also like to thank our colleagues at the Centers for Disease Control for their efforts and support that made this resource available. This research is sponsored by the National Science Foundation (Award #0844032), Dell, Intel, Microsoft, and The MathWorks. We would also like to thank our colleagues at the Centers for Disease Control for their efforts and support that made this resource available.