Research Plan: Enabling MPI on the LHC Computing Grid

Research Plan for RP2
University of Amsterdam
MSc in System and Network Engineering

Class of 2005-2006

Richard de Jong, Matthijs Koot
{rjong,mrkoot}@os3.nl

June 7, 2006

Contents

1 Introduction

As part of our Master of Science study in the field of System and Network Engineering, at the University of Amsterdam, we will be doing research at CERN [1], Geneva, on enabling MPI-based parallel computing on the LHC Computing Grid (LCG). The research is (formally) performed on behalf of NIKHEF [2], Amsterdam, and is supervised by Louis Poncet at CERN.

2 Research goal

TBD

Parallel programming is the art of using multiple processors to solve a single problem. The traditional paradigm of computer architecture is the opposite: to solve multiple problems with a single processor (serial computing). With millions of scalar computers now connected through the Internet, the perspective on computing is slowly changing to that of 'the network is the computer'; in stead of using a single processor to solve a problem, a global network of processors is now available to solve (the same or larger) problems. Historic methods of parallel computing included threading, IPC and Parallel Virtual Machines (*PVM*). None of them, however, are really suitable for a 'heavily distributed' environment spanning the globe [9]. In response to this, a new protocol was designed as a successor to PVM and has now, a decade later, become a *de facto* standard for massively parallel computing: the Message Passing Interface, or *MPI* [5]. MPI, which is a library specification, solves the problem of interprocess and job communication and allows data to be passed between processes in a distributed memory environment. The prime quality attributes of MPI are source code portability, i.e. to support heterogeneous parallel architectures, and to allow efficient implementation, i.e. allow optimization for certain hardware platform [11]. Although MPI may be used in shared-memory architectures (SMP, NUMA), it's original design was focussed at distributed-memory architecture. It is the latter type of environment in which MPI is used at the LCG; many scalar processors with their own memory, mostly grouped into scientific computing clusters, connected over the Internet.

So far for MPI and parallel computing. Where does the *grid* fit in? According to Ian Foster, 'father of the grid', a grid is a system that [6]:

1. ... coordinates resources that are not subject to centralized control;
2. ... using standard, open, general-purpose protocols and interfaces;
3. ... to deliver nontrivial qualities of service.

Thus, it is more than simply a bunch of interconnected computational resources (which might rather be called a *cluster*). The most prevalent vision on grid computing is that of 'service-oriented computing' [7, 8], i.e., considering computational resources to be like water and electricity facilities. Exploiting these

means for scientific purposes is also denoted with the term *e-Science*, as coined by John Taylor [7]. Building on knowledge and code from Globus Toolkit, the European DataGrid (EDG) project, the Enabling Grids for E-sciencE (EGEE) project, the LCG project at CERN aims to deliver a production-quality world-wide grid for scientific purposes (and perhaps commercial usage at a later stage). A multilayer model has been developed for grid architectures. This model, which is semantically comparable with the TCP/IP model, is depicted in figure 1. It is believed that like any system to be connected to the Internet needs an IP-address, any system to be connected to the Grid shall need to support the Open Grid Services Architecture. The functions placed between the *User Applications* layer and the *Fabric* layer are provided by *Grid middleware*. In the case of the LCG, that middleware is called *gLite* (spin-off credited to the EGEE project). This middleware is typically deployed through Yet Another Installation Method, or *YAIM*.

— *BOF goals (TBD)* — The first goal of our research is to integrate (and demonstrate) MPI into the gLite middleware and the YAIM deployment scheme, so that, if enabled during install, MPI jobs may be submitted through the LCG grid interface. The second goal is to evaluate YAIM by assessing it on various quality attributes, e.g. extensibility. The third goal is to report on the scalability and dependency aspects of the MPI library with respect to the current grid middleware, as well as the modifications needed to support MPI. — *EOF goals (TBD)* —

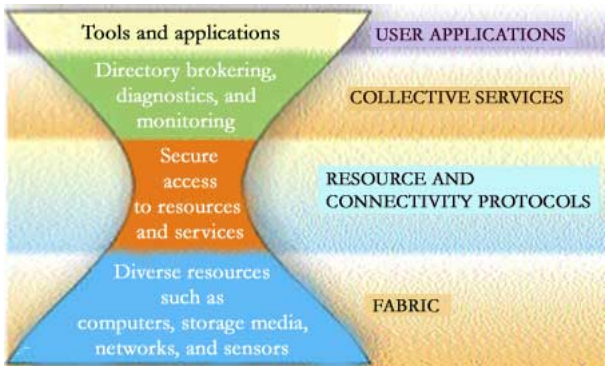

Figure 1: The layered grid-model

3 Project organization

The project is organized in an informal way and consists of the people and roles defined below. As a means of minimum quality assurance, we will have a weekly feedback meeting at CERNs site in Meyrin. The meetings will be scheduled as time goes by.

Name	Role(s)	Contact
Richard de Jong	project member	rjong at os3.nl
Matthijs Koot	project member	mrkoot at os3.nl
David Groep	supervisor NIKHEF	david.groep at nikhef.nl
Jeff Templon	contact at NIKHEF	jeff.templon at nikhef.nl
Louis Poncet	supervisor CERN	louis.poncet at cern.ch
Cees de Laat	university supervisor	delaat at science.uva.nl

4 Project resources

Project resources may be roughly divided in two categories: human and non-human resources. For this project, no human resources other than those who are part of the project organization are anticipated. Concerning the non-human category, the following resources are anticipated:

MoSCoW	Resource
MUST	Access to the LCG.
SHOULD	Access to gLite architecture documents.

TBD: describe what we (don't) have

5 Project deliverables

The deliverables for this project are as follows:

- research plan (this document);
- research report;
- production-ready integration of MPI within gLite (RPM package).

TBD: describe each deliverable

6 Project planning

Aligning with the requirements of our university, these dates are considered the deadlines for this project:

Deadline	**Deliverable**	**Comments**
June 9th, 2006	Research Plan	Will be sent to David Groep, Louis Poncet and Cees de Laat for approval.
June 30th, 2006	Research Report	Will be sent to David Groep, Louis Poncet and Cees de Laat for approval.
July 8th, 2006	Presentation	A public event in the Turingzaal at SARA, Amsterdam.

During the period of June 5th to June 9th, we will be gaining more in-depth knowledge on grid computing and MPI by reading related work, tutorials and performing some simple MPI jobs on the LCG. From there, we should be able to understand the issues for which requirements will be defined. Between June 12th and June 16th, we will be working on a gLite-based package for OpenMPI, as well as how to integrate it into YAIM. In the week of June 19th to June 23rd, we will investigate the quality of YAIM and advise on how it might be improved. The latter may include best practices defined by the Global Grid Forum, as well as general software quality attributes. In the last week of June 26th to June 30th we will deliver our research report to the those listed in the above table. We anticipate a buffer of three days in the last week for delays in our research.

7 Copyrights and ownership

All documents which are created as a part of this project will be licensed under the Creative Commons 2.5 Attribute license [3]. All source and object code which is produced as a part of this project will be licensed under the revised BSD license [4].

Appendix A - Structure of the Research Report

The research report, our main deliverable, will consist of the elements summarized in the next subsections. As our research progresses, so will the contents of these sections.

Introduction to Enabling MPI on the LHC Computing Grid

This section will contain a general introduction into grid computing and parallel programming with MPI.

Scope

This section will contain the definite specification of the scope of our research. The scope may be subject to slight changes at the beginning of our project due to knowledge obtained during that period. Any such changes will be definite *only* after approval of the supervisor(s).

Related work

TBD

Section++

Future work

This section will contain some suggestions for future work.

Conclusion

The report will be finalized with a conclusion summarizing our results and how they compare with the goals defined in the research plan.

Appendix B - The BSD license for this project

Copyright (c) 2006, Richard de Jong and Matthijs Koot
All rights reserved.

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

* Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
* Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.
* Neither the name of the University of Amsterdam nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

References

[1] European Organization for Nuclear Research, www.cern.ch

[2] NIKHEF: Dutch Institute for High Energy and Nuclear Physics, www.nikhef.nl

[3] Creative Commons: Creative Commons Attribution 2.5 license, www.creativecommons.org

[4] Open Source Initiative, The BSD Licensewww.opensource.org

[5] MPI Forum: Message Passing Interface (MPI) Forum homepage, 1998, http://www.mpi-forum.org/docs/docs.html

[6] Foster, Ian: *"What is the Grid? A Three Point Checklist."*http://www-fp.mcs.anl.gov/ foster/Articles/WhatIsTheGrid.pdf

[7] Foster, Ian: *"Service-Oriented Science."*, 2005, http://www.sciencemag.org/cgi/content/short/308/5723/814

[8] Papazoglou, M. and Georgakopoulos, D.: *"Service-oriented computing: Introduction"*, http://portal.acm.org/citation.cfm?doid=944217.944233

[9] Gropp, William and Lusk, Ewing: *"PVM and MPI are completely different"*, 1997, http://citeseer.ist.psu.edu/573977.html

[10] Gropp, William and Lusk, Ewing: *"Goals Guiding Design: PVM and MPI"*, 2002, http://citeseer.ist.psu.edu/568858.html

[11] Fagg, Graham and London, Kevin: *"MPI Inter-connection and Control"*, 1998, http://citeseer.ist.psu.edu/400213.html