DESIGN AND ANALYSIS OF A COMMUNICATION MIDDLEWARE FOR MULTIPLE NETWORK INTERFACES

By
Nader Mohamed

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Computer Science

Under the Supervision of Professor Hong Jiang

Lincoln, Nebraska

June 2004

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

UMI®

UMI Microform 3147147

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

DISSERTATION TITLE

Design and Analysis of a Communication Middleware for Multiple Network Interfaces

BY

Nader Mohamed

SUPERVISORY COMMITTEE:

	Approved	Date
Signature		6/23/04
Typed Name	Dr. Hong Jiang	
Signature		6/23/04
Typed Name	Dr. Steve Goddard	
Signature		6/23/04
Typed Name	Dr. Stephen Scott	
Signature		6/23/04
Typed Name	Dr. Sharad Seth	
Signature		6/23/04
Typed Name	Dr. Xiao Cheng Zeng	
Signature		
Typed Name		

University of Nebraska | GRADUATE COLLEGE

DESIGN AND ANALYSIS OF A COMMUNICATION MIDDLEWARE FOR MULTIPLE NETWORK INTERFACES

Nader Mohamed, Ph.D.
University of Nebraska-Lincoln, 2004

Adviser: Hong Jiang

Effectively utilizing multiple network interfaces and networks can enhance end-to-end communication performance in local area networks (LAN), system area networks (SAN), and wide-area-networks (WAN). Multiple homogeneous or heterogeneous network interfaces connected to single or multiple homogeneous or heterogeneous networks may exist within and/or among some homogenous or heterogeneous systems. However, current network protocols and services cannot seamlessly and effectively utilize the existing multiple network interfaces. In this dissertation, two communication middleware models that seamlessly and efficiently utilize multiple network interfaces are designed and analyzed. The first model is Multiple-Network-Interface Socket (MuniSocket). MuniSocket provides parallel data transfer over multiple network interfaces. The second model is Multiple-Network-Interface Socket for Clusters (MuniCluster). MuniCluster provides configurable, flexible, and expandable communication infrastructure specifically for cluster networks. These models are configurable to deal with a variety of interconnection structural and traffic scenarios. In addition, they transparently provide an expandable high-bandwidth solution that reduces message transfer time and facilitate dynamic load balancing among the underlying multiple networks. The middleware approach of utilizing multiple network interfaces has

the unique advantages of providing flexibility in handling application demands and being independent of the hardware and protocols used in the lower layers. In addition, it is highly portable thus it facilitates handling heterogeneous systems, platforms, networks, and network interfaces. The main contributions of this research are: (1) the design of a configurable communication middleware for utilizing multiple network interfaces to enhance end-to-end communication performance, MuniSocket; (2) the development of reliable parallel transfer mechanisms on top of the reliable and unreliable transport services; (3) the study of the scalability problems in bulk data transfer in wide area networks for Grid applications and a demonstration on how the middleware model can solve most of these problems; (4) the development of different optimization techniques for multiple network interface communication configurations; and (5) the optimization of the proposed model for cluster environments, MuniCluster.

This is for my sons Hashim and Qassim.

Acknowledgement

To Jameela, (my undergraduate, master, and Ph.D. colleague; my best friend; my research partner; my wife; the mother of my boys Hashim and Qassim; and my colleague in my future academic career). She has been supportive and encouraging all the way, for which I will always be grateful. To my family including my Father Faisal, my mother Taj, my brothers Reyad, Fuad, and Eyad, and my sisters Dr. Batool, Dr. Lamya, and Sawsan for all their encouragement and support in pursuing my graduate studies. To Dr. Hong Jiang, my supervisor, who provides invaluable and constructive guidance, support, and positive comments. I am grateful for the tremendous time and effort he provides. I would like thank Dr. David Swanson for his support, assistance, and help in this research. In addition, I am also thankful to my Ph.D. supervisory committee members Dr. Steve Goddard (dissertation reader), Dr. Stephen Scott (dissertation reader), Dr. Xiao Cheng Zeng, and Dr. Sharad Seth for their support and valuable comments and suggestions. To Donna McCarthy for all her help and support during my Ph.D. studies. I would also like to thank the members of the secure distributed information (SDI) group and the research computing facility (RCF) at the University of Nebraska-Lincoln for their continuous help and support. I would like to thank Sumanth Jannyavula-Venk for all his help in setting the experimental environments. This research was partially supported by a National Science Foundation grant (EPS-0091900) and a Nebraska University Foundation grant (26-0511-0019), for which I am very grateful.

Table of Contents

List of Figures

List of Tables

The Dissertation Publications

[1] N. Mohamed, J. Al-Jaroodi, H. Jiang, and D. Swanson, "High-Performance Message Striping over Reliable Transport Protocols," accepted to appear in *Journal of Supercomputing*, Special Issue on Infrastructures and Applications for Cluster and Grid Computing Environments, 2004.

[2] N. Mohamed, J. Al-Jaroodi, and H. Jiang, "Configurable Communication Middleware for Clusters with Multiple Interconnections," accepted to appear in *IEICE Transactions of Information and Systems*, Special Issue on Hardware/Software Support for High Performance Scientific and Engineering Computing, Guest editors: Minyi Guo and Laurence T. Yang, July 2004.

[3] N. Mohamed, J. Al-Jaroodi, and H. Jiang, "Flexible Mechanisms for Performance Enhancements of Cluster Networks," in *proceedings of The 23rd IEEE International Performance, Computing, and Communications Conference* (IPCCC 2004), Phoenix, Arizona, pp. 359-366, April 2004.

[4] N. Mohamed, J. Al-Jaroodi, and H. Jiang, "Extensible Communication Architecture for Grid Nodes," in *proceedings of The 5th International Conference on Information Technology: Coding and Computing* (ITCC 2004), Track on Methodologies, Technologies, and Applications in Distributed and Grid Systems, Las Vegas, Nevada, IEEE, volume 2, pp. 40-44, April 2004.

[5] N. Mohamed, J. Al-Jaroodi, H. Jiang, and D. Swanson, "Reliable Bulk Data Transfer Protocols over Multiple Network Interfaces," in *proceedings of The 3rd*

IEEE International Conference on Networking (ICN'04), Guadeloupe, French Caribbean, 8 pages, March 2004.

[6] N. Mohamed, J. Al-Jaroodi, H. Jiang, and D. Swanson, "Scalable Bulk Data Transfer in Wide Area Networks," in *International Journal of High Performance Computing Applications* – Special Issue on Grid Computing: Infrastructure and Applications, Guest editor: David Walker, Volume 17, No. 3, pp. 237-248, August 2003.

[7] N. Mohamed, J. Al-Jaroodi, H. Jiang, and D. Swanson, "A Middleware-Level Parallel Transfer Technique over Multiple Network Interfaces," in *proceedings of The ClusterWorld Conference and Expo*, San Jose, California, CD-ROM – 15 pages, June 2003.

[8] N. Mohamed, J. Al-Jaroodi, H. Jiang, and D. Swanson, "Performance Properties of Combined Heterogeneous Networks," in *proceedings of the IEEE IPDPS 2003, IEEE/ACM International Workshop on Performance Modeling, Evaluation, and Optimization of Parallel and Distributed Systems* (PMEO-PDS'03), Nice, France, IEEE, CD-ROM – 7 pages, April 2003.

[9] N. Mohamed, J. Al-Jaroodi, H. Jiang, and D. Swanson, "A User-Level Socket Layer Over Multiple Physical Network Interfaces," in *proceedings of The 14th International Conference on Parallel and Distributed Computing and Systems* – High Performance Computing and Networking Symposium, Cambridge, Massachusetts, pp. 810-815, November 2002.

Chapter 1. Introduction

Applications such as network-based multimedia storage, remote satellite observations, distributed data mining, distributed scientific simulations, and distributed geographic information systems are both *compute intensive*, requiring scalable high-processing power, and *data intensive*, demanding reliable, scalable high-bandwidth communication infrastructure for high-volume and high-speed data access. Currently, clusters of PCs/workstations and grids, owing to their significant cost-effectiveness, are poised to provide the most suitable infrastructure for these applications. Such systems provide processing power comparable to special purpose high-end multi-processor systems at a fraction of the cost.

One of the main challenges in developing infrastructure services for Cluster and Grid computing is the heterogeneity in machine architectures, operating systems, and network resources. This heterogeneity forces the services to be implemented at the middleware-level so that they will be easily ported to and utilized by different types of systems. Implementing tools and services that need special kernel functions results in poor portability and utilization of such tools and services. Current Cluster and Grid tools and services have different degrees of flexibility in dealing with heterogeneity in systems and networks. Another important issue in Cluster and Grid computing is the scalability of services. Grid computing aims to utilize available distributed software and hardware resources on a large scale spanning the whole nation and even the globe. This requires the Cluster and Grid services to be scalable to efficiently utilize the available resources.

One solution for expandability of processing power and storage is to add more nodes and storage units to the cluster or grid. Similarly, adding more network interfaces and connections can increase the total communication bandwidth among the cluster nodes and grid components. However, current network protocols, software, and APIs such as Sockets are designed for a single physical network interface. This gives rise to the need for protocols, and software services that can support multiple physical network interfaces on each node and provide transparent and efficient utilization of these interfaces.

In this dissertation, we study the design of two scalable and flexible communication middleware models that exploit multiple network interfaces to enhance the overall communication performance in networks such as local area networks (LAN), system area networks (SAN), and wide-area-networks (WAN). The first model is called Multiple-Network-Interface Socket (MuniSocket) and the second is called Multiple-Network-Interface Socket for Clusters (MuniCluster). These models have the flexibility to deal with homogeneous and heterogeneous systems, networks, and interfaces. This research is motivated by the diversity in interconnection network structure and traffic scenarios that exists in cluster and grid environments. In addition, we investigate a number of possible optimization techniques for these models.

In the rest of this dissertation, we will motivate the research by reviewing the current problems of local area networks and cluster interconnections and describing the need for solutions to these problems in Chapter 2. In addition, the problem statements are presented in Chapter 2.

In Chapter 3, we introduce the user-level socket model that utilizes multiple physical network connections on a cluster or among grid components, transparent to the user applications. In this model, user messages are fragmented into uniform packets of predetermined size. Then the fragments are numbered and transferred in parallel through multiple network interfaces, via one or more physical networks to the destination, where the fragments are reconstructed back to the original message, transparently to the user application. The Multiple Network Interface Socket (MuniSocket) prototype has been implemented based on this model. The main difference between MuniSocket and the standard Socket is that MuniSocket processes and transfers large user messages in parallel, fully utilizing the existing multiple-network interconnects, while the standard Socket processes and transfers messages sequentially through a single network interface. In other words, MuniSocket has the potential of providing expandable bandwidth, load balancing, and fault-tolerance for machines connected by multiple interconnection networks. This model is based on unreliable transport protocols, typified by UDP (User Datagram Protocol). In addition, in this chapter we design a reliable parallel transfer mechanism on top of the unreliable transport protocols.

Chapter 4 discusses an application domain for UDP-Based MuniSocket that supports bulk data transfer for Grid computing on WAN, along with some experimental results. The main contributions of this chapter are (1) the identification of the scalability limitations of increasing bandwidth among Data Grid components, (2) the classification of the approaches for increasing bandwidth among Data Grid components on WAN, and (3) the performance evaluation of UDP-Based MuniSocket on WAN. In this chapter, we further extend the MuniSocket model to provide scalable high-bandwidth communication

for Grid applications over WAN. We also evaluate the performance of MuniSocket over a simulated WAN to show its operations and performance on Grids with heterogeneous systems and network resources.

Chapter 5 further discusses the MuniSocket model through the design of a TCP-Based MuniSocket model based on reliable transport protocols (e.g. TCP). In addition, we discuss a number of potential optimizations to enhance the TCP-Based MuniSocket's performance. We extend our discussion of the high-performance, middleware-level concurrent message striping technique in the context of reliable transfer protocols. This striping technique provides a scalable network bandwidth and a portable solution for data transfer among heterogeneous systems. It reduces message transfer time and facilitates dynamic load balancing among the underlying multiple networks. In addition, we introduce some techniques to enhance the performance of this middleware-level concurrent striping technique of message transfer over multiple networks. These techniques reduce the contention on the striping counter and the overhead of sending sequence numbers (additional header) with each fragment of the striped messages. These techniques rely on the features available in reliable transport protocols, such as in-order and guaranteed delivery of packets, to significantly reduce some of the striping overhead. We also introduce a technique to enhance communication dependability in the dual network case.

Chapter 6 describes MuniCluster, an extended model of MuniSocket to enhance the performance properties and flexibility in cluster networks. In clusters, applications benefit differently from communication performance enhancements, depending on their reliance on different properties at varying situations. For example, bulk data transfer

requires high bandwidth, while multi-node graphical visualization requires low latency and high throughput among participating nodes. However, most cluster communication software and hardware infrastructures do not deploy flexible mechanisms to utilize multiple networks and network interfaces to accommodate these differences in performance requirements. The main contributions of Chapter 6 are (1) the identification and classification of possible end-to-end enhancements and their flexibility requirements for cluster network properties, and (2) the development of flexible mechanisms to enhance the cluster network properties. The mechanisms provide scalable cluster communication by utilizing multiple interface cards connected by single or multiple networks. These mechanisms include parallel transfer, separation of transfer channels for small messages and large messages, and load balancing between multiple network channels. Furthermore, the introduced techniques allow for easy and flexible ways to enable heterogeneous clusters with varying capacities and number of network interfaces to utilize all of these resources efficiently.

Chapter 7 covers related work and their advantages and disadvantages, along with some discussions. And finally, Chapter 8 concludes the dissertation with a summary of our contributions, the observed advantages, and the proposed future work.

Chapter 2. Motivation, Problem Statement, and Scope of the Dissertation Research

This chapter discusses the main motivation for this dissertation research and the problem statement in Section 2.1. In this section we review some of the current technologies, distributed applications demands, the current problems, and the existing approaches to enhance communication properties. Then Section 2.2 defines the scope of research by outlining our approaches to providing flexible and configurable solutions to enhance the communications properties.

2.1 Motivation and Problem Statement

Although current technologies have made it rather simple to acquire high performance components at reasonable costs, the available software technologies are not yet adequate to seamlessly handle the variety of possible configurations of hardware components. For example, the configurations of most cluster systems include two or more network interface cards (NIC) per node, but the applications cannot seamlessly utilize these NICs simultaneously. Conventionally, each NIC on a machine is assigned an IP address, forcing each machine to have multiple IP addresses. Currently, most of the distributed applications use TCP/IP protocols, thus requiring the use of a fixed IP address or a domain name in environments that are equipped with a domain name server (DNS) to identify the machine. This leads to the necessary requirement that a given application use a single IP address on each node to establish connections with other nodes, hence preventing the application from utilizing other available networks. This means that the bandwidth available for communication to any application is bounded by the bandwidth

available on the single network interface associated with the assigned IP address. Thus, even if there are multiple networks connecting the nodes of the cluster, only one of them will be utilized for any given application at any point in time. To satisfy the application's demand for higher bandwidth with the current technology, the existing networks have to be replaced by more advanced/faster networks, which is a costly and non-scalable solution. This problem exists in local area networks. Table 2-1 shows the different possible scenarios in local area networks in which machines are connected by multiple network interfaces or/and networks. The machines can be homogeneous or heterogeneous in architectures, capacities, and operating systems. The network interfaces can be homogeneous or heterogeneous as well. The network interfaces can be connected to single or multiple networks. Another problem that faces applications on clusters is the lack of information available to each application about the NIC utilization by the other applications. Current settings require each application to bind the IP addresses it utilizes independent from other applications, thus multiple applications on a node cannot coordinate their network utilization. For example, it could easily happen that one application starts using one network and then the next application binds its sockets to the same network even though the second network is free.

Many research laboratories and organizations have multiple high-performance heterogeneous machines for different application purposes. The heterogeneity of these machines usually lies in their differences in architectures, operating systems, processing power, memory, storage, and communication capacities. There is an increasing interest in utilizing the hardware and software resources available among a collection of existing heterogeneous systems to implement high-performance, scalable clusters capable of