

*Commenced Publication in 1973*

Founding and Former Series Editors:

Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

## Editorial Board

David Hutchison

*Lancaster University, UK*

Takeo Kanade

*Carnegie Mellon University, Pittsburgh, PA, USA*

Josef Kittler

*University of Surrey, Guildford, UK*

Jon M. Kleinberg

*Cornell University, Ithaca, NY, USA*

Alfred Kobsa

*University of California, Irvine, CA, USA*

Friedemann Mattern

*ETH Zurich, Switzerland*

John C. Mitchell

*Stanford University, CA, USA*

Moni Naor

*Weizmann Institute of Science, Rehovot, Israel*

Oscar Nierstrasz

*University of Bern, Switzerland*

C. Pandu Rangan

*Indian Institute of Technology, Madras, India*

Bernhard Steffen

*TU Dortmund University, Germany*

Madhu Sudan

*Microsoft Research, Cambridge, MA, USA*

Demetri Terzopoulos

*University of California, Los Angeles, CA, USA*

Doug Tygar

*University of California, Berkeley, CA, USA*

Gerhard Weikum

*Max Planck Institute for Informatics, Saarbruecken, Germany*

Jesper Larsson Träff Siegfried Benkner  
Jack Dongarra (Eds.)

# Recent Advances in the Message Passing Interface

19th European MPI Users' Group Meeting  
EuroMPI 2012  
Vienna, Austria, September 23-26, 2012  
Proceedings

## Volume Editors

Jesper Larsson Träff  
Vienna University of Technology  
Faculty of Informatics  
Institute of Information Systems  
Research Group Parallel Computing  
Favoritenstr. 16  
1040 Vienna, Austria  
E-mail: traff@par.tuwien.ac.at

Siegfried Benkner  
University of Vienna  
Faculty of Computer Science  
Research Group Scientific Computing  
Währinger Str. 29/6.21  
1090 Vienna, Austria  
E-mail: siegfried.benkner@univie.ac.at

Jack Dongarra  
University of Tennessee  
Department of Electrical Engineering  
and Computer Science  
Knoxville, TN 37996, USA  
E-mail: dongarra@cs.utk.edu

ISSN 0302-9743	e-ISSN 1611-3349
ISBN 978-3-642-33517-4	e-ISBN 978-3-642-33518-1
DOI 10.1007/978-3-642-33518-1	
Springer Heidelberg Dordrecht London New York	

Library of Congress Control Number: Applied for

CR Subject Classification (1998): C.2.4, F.2, D.2, C.2, H.4, D.4

LNCS Sublibrary: SL 2 – Programming and Software Engineering

© Springer-Verlag Berlin Heidelberg 2012

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

*Typesetting:* Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

# Preface

Extended message-passing style parallel programming with MPI remains the most important and successful paradigm for programming hybrid, distributed memory parallel systems and achieving high application efficiency. MPI, the Message-Passing Interface, introduced more than 20 years ago, has been an extremely efficient and productive interface (both in qualitative and quantitative terms), and proven surprisingly robust in the face of very radical changes in systems configurations, capabilities, and scale over the past decades. This has entailed an immense amount of work, both in improvement of the implementations of MPI, mostly done by research labs and in academic environments, but also in part by commercial vendors (that often base their developments on open implementations from labs and academe), and in exploration and extension of the standard itself often as driven by application needs. The EuroMPI conference series has provided and will continue to provide an important forum for MPI developers, researchers in message-passing parallel programming, application developers, users, and students to meet and discuss specific issues related to MPI; always with a look towards new trends and developments of related or alternative interfaces for high-performance parallel programming, and often in quite close interaction with important HPC vendors. In the past five years the MPI Forum has been active in revising and extending the MPI standard, addressing among others issues of scalability, and has brought out consolidated versions of MPI 2, as well as drafts for more significant extensions to go into an upcoming MPI 3.0 version of the standard. In this process EuroMPI has played a role in testing new proposals for MPI 3, for example on fault-tolerance, collective communication, interaction with threads, and other matters. EuroMPI is one of the few meetings where these kinds of specific explorations related to the concrete MPI standard can be discussed, and should be used also in the future for more such research. It is a community conviction that other paradigms and interfaces for highly parallel distributed memory programming must do as well as MPI in order to be successful, and that there is consequentially much to learn from MPI and ongoing research activities as presented at the EuroMPI conference.

EuroMPI 2012 featured 22 technical presentations on MPI implementation techniques and issues, benchmarking and performance analysis, programming models and new architectures, run-time support, fault-tolerance, message-passing algorithms, and applications. A special session on Improving MPI User and Developer Interaction (IMUDI), introduced with EuroMPI 2011, was dedicated to intensifying interaction between users and implementors of MPI, in particular to

make user expectations and desiderata regarding the standard (and its implementations) explicit. The conference also featured four invited talks on MPI 3 and beyond (Gropp), the Fujitsu petaflop K computer and its MPI (Sumimoto), the impact of MPI on design of efficient interconnect hardware (Brüning), and the prospects of applying advanced compiler optimizations to MPI programs (Danalis), as well as two tutorials on advanced MPI and performance engineering. The conference was rounded off with a vendor session, a report from the MPI Forum, discussion slots, and a poster exhibition. Papers and abstracts can be found on the following pages. The meeting program and (most of) the talks can be found at [www.eurompi2012.org](http://www.eurompi2012.org).

EuroMPI is the successor to the EuroPVM/MPI user group meeting series (since 2010), making EuroMPI 2012 the 19th event of this kind. EuroMPI takes place each year at a different European location; the 2012 meeting was held in Vienna, Austria, organized jointly by Vienna University of Technology (TU Wien) and the University of Vienna. Previous meetings were held in Santorini (2011), Stuttgart (2010), Espoo (2009), Dublin (2008), Paris (2007), Bonn (2006), Sorrento (2005), Budapest (2004), Venice (2003), Linz (2002), Santorini (2001), Balatonfüred (2000), Barcelona (1999), Liverpool (1998), Cracow (1997), Munich (1996), Lyon (1995), and Rome (1994). The meeting took place at the Austrian Academy of Sciences, during September 23–26, 2012.

In reaction to the call for papers that was first published late 2011, we received a total of 47 submissions by the (extended) submission deadline on May 16th, clearly fewer than hoped for. The low number of submissions possibly reflects the universally more difficult funding situation for conference travel. EuroMPI has so far had a very good record with respect to attendance and presentation with as good as no no-shows; potential contributors who knew in advance that they might not be able to travel may have chosen to submit to geographically closer forums. It might also reflect the (positive) fact that good MPI work, whether in implementations or applications, can also be presented at broader parallel processing conferences. All 47 submissions were in scope, and were reviewed by program committee members (with only relatively few external referees) with each paper getting between 3 and 5 reviews. An effort was made to provide informative and helpful feedback to authors. Based on the reviews, the program chairs selected 22 submissions as regular papers, and 7 papers for presentation as posters. Regular papers were allotted 10 pages in the proceedings, and a 30 minute slot for presentation. Among the regular papers, a handful of the strongest and best presented are invited for a Special Issue of the Springer “Computing” journal. These extended papers will again be reviewed by members of the EuroMPI 2012 program committee as well as by new external reviewers.

The program chairs and general chair would like to thank all authors who submitted their contributions to EuroMPI 2012; the program committee members for their work in getting the submissions reviewed, mostly in time and with good-quality, informative reviews; our sponsors who contributed significantly toward making the conference feasible; and all who attended the meeting in Vienna. We hope that the EuroMPI 2012 conference had something to offer for all, and will remain a solid forum for high-quality MPI-related work as it goes into its third decade.

September 2012



Jesper Larsson Träff  
Siegfried Benkner  
Jack Dongarra

# Organization

EuroMPI 2012 was organized jointly by Vienna University of Technology (TU Wien) and the University of Vienna, in association with the Innovative Computing Laboratory of the University of Tennessee.

## General Chair

Jack Dongarra	University of Tennessee, USA
---------------	------------------------------

## Program Chairs

Siegfried Benkner	University of Vienna, Austria
Jesper Larsson Träff	Vienna University of Technology, Austria

## Program Committee

Pavan Balaji	Argonne National Laboratory, USA
Siegfried Benkner	University of Vienna, Austria
Gil Bloch	Mellanox Technologies, Israel
George Bosilca	University of Tennessee, Knoxville, USA
Ron Brightwell	Sandia National Laboratories, Albuquerque, USA
Darius Buntinas	Argonne National Laboratory, USA
Franck Cappello	INRIA, France and University of Illinois at Urbana-Champaign, USA
Gilles Civario	Irish Centre for High-End Computing, Ireland
Yiannis Cotronis	University of Athens, Greece
Jim Cownie	Intel, UK
Anthony Danalis	University of Tennessee, Knoxville, USA
Bronis R. de Supinski	Lawrence Livermore National Laboratory, USA
Luiz DeRose	Cray, USA
Edgar Gabriel	University of Houston, USA
Brice Goglin	INRIA, Bordeaux, France
David Goodell	Argonne National Laboratory, USA
Ganesh Gopalakrishnan	University of Utah, USA
Richard Graham	Oak Ridge National Laboratory, USA
William Gropp	University of Illinois at Urbana-Champaign, USA
Thomas Herault	University of Tennessee, Knoxville, USA
Torsten Hoefer	University of Illinois at Urbana-Champaign, USA
Yutaka Ishikawa	University of Tokyo, Japan

Michael Kagan	Mellanox Technologies, Israel
Rainer Keller	HFT Stuttgart, University of Applied Science, Germany
Dries Kimpe	Argonne National Laboratory, USA
Jesus Labarta	Technical University of Catalonia, Barcelona Supercomputing Center, Spain
Dong Li	Oak Ridge National Laboratory, USA
Ewing Rusty Lusk	Argonne National Laboratory, USA
Amith Rajith Mamidala	IBM, USA
Satoshi Matsuoka	Tokyo Institute of Technology, Japan
Guillaume Mercier	INRIA, France
Bernd Mohr	Jülich Supercomputing Centre, Germany
Matthias Mueller	TU Dresden, Germany
Rolf Rabenseifner	High Performance Computing Center Stuttgart (HLRS), Germany
Thomas Rauber	University of Bayreuth, Germany
Rolf Riesen	IBM, Ireland
Robert Ross	Argonne National Laboratory, USA
Peter Sanders	Karlsruhe Institute of Technology, Germany
Mitsuhisa Sato	University of Tsukuba, Japan
Saba Sehrish	Northwestern University, USA
Christian Siebert	University of Aachen, Germany
Stephen Siegel	University of Delaware, USA
Anna Sikora	Autonomous University of Barcelona, Spain
Jeff Squyres	Cisco, USA
Shinji Sumimoto	Fujitsu Ltd., Japan
Rajeev Thakur	Argonne National Laboratory, USA
Vinod Tipparaju	AMD, USA
Carsten Trinitis	Technical University of Munich, Germany
Denis Trystram	Grenoble Institute of Technology, France
Jesper Larsson Träff	Vienna University of Technology, Austria
Keith Underwood	Intel, USA
Robert Van De Geijn	The University of Texas at Austin, USA
Alan Wagner	University of British Columbia, Canada
Roland Wismüller	University of Siegen, Germany
Xin Yuan	Florida State University, USA

## External Referees

Sriram	Sascha Hunold	Hitoshi Sato
Aananthakrishnan	Benny Koren	Subodh Sharma
Leonardo Bautista	Guodong Li	Keita Teranishi
Eduardo Cesar	Grant Mackey	Francois Trahay
Wei-Fan Chiang	Sabri Pllana	
Aleksandr Drozd	Claudia Rosas	



## Local Organization

Jesper Larsson Träff, Vienna University of Technology  
Siegfried Benkner, University of Vienna

Enes Bajrovic, University of Vienna  
Christine Kamper, Vienna University of Technology  
Margret Steinbuch, Vienna University of Technology  
Angelika Wiesinger, University of Vienna

## Sponsors

The conference would not have been possible without financial support from sponsors, and we therefore gratefully acknowledge the support and contribution of this years' sponsors to a successful meeting. Platinum and Gold sponsors also contributed with technically oriented talks in the vendor session, an important part of the conference for getting technically oriented information from relevant HPC and interconnect vendors and software developers.

### Platinum level sponsor



### Gold level sponsors



### Silver level sponsors



# Table of Contents

## Invited Talks

MPI 3 and Beyond: Why MPI Is Successful and What Challenges It Faces . . . . .	1
<i>William Gropp</i>	
MPI Functions and Their Impact on Interconnect Hardware . . . . .	10
<i>Ulrich Brüning</i>	
The MPI Communication Library for the K Computer: Its Design and Implementation . . . . .	11
<i>Shinji Sumimoto</i>	
MPI and Compiler Technology: A Love-Hate Relationship . . . . .	12
<i>Anthony Danalis</i>	

## Tutorials

Advanced MPI Including New MPI-3 Features . . . . .	14
<i>William Gropp, Ewing Lusk, and Rajeev Thakur</i>	
Hands-on Practical Hybrid Parallel Application Performance Engineering . . . . .	15
<i>Markus Geimer, Michael Gerndt, Sameer Shende, Bert Wesarg, and Brian Wylie</i>	

## MPI Implementation Techniques and Issues

Adaptive Strategy for One-Sided Communication in MPICH2 . . . . .	16
<i>Xin Zhao, Gopalakrishnan Santhanaraman, and William Gropp</i>	
A Low Impact Flow Control Implementation for Offload Communication Interfaces . . . . .	27
<i>Brian W. Barrett, Ron Brightwell, and Keith D. Underwood</i>	
Improving MPI Communication Overlap with Collaborative Polling . . .	37
<i>Sylvain Didelot, Patrick Carribault, Marc Pérache, and William Jalby</i>	
Delegation-Based MPI Communications for a Hybrid Parallel Computer with Many-Core Architecture . . . . .	47
<i>Kazumi Yoshinaga, Yuichi Tsujita, Atsushi Hori, Mikiko Sato, Mitaro Namiki, and Yutaka Ishikawa</i>	

Efficient Multithreaded Context ID Allocation in MPI . . . . .	57
<i>James Dinan, David Goodell, William Gropp, Rajeev Thakur, and Pavan Balaji</i>	
Collectives on Two-Tier Direct Networks . . . . .	67
<i>Nikhil Jain, JohnMark Lau, and Laxmikant Kale</i>	
Exploiting Atomic Operations for Barrier on Cray XE/XK Systems . . . .	78
<i>Manjunath Gorentla Venkata, Richard L. Graham, Joshua S. Ladd, Pavel Shamis, Nathan T. Hjelm, and Samuel K. Gutierrez</i>	
Exact Dependence Analysis for Increased Communication Overlap . . . . .	89
<i>Simone Pellegrini, Torsten Hoeﬂer, and Thomas Fahringer</i>	

## Benchmarking and Performance Analysis

mpicroscope: Towards an MPI Benchmark Tool for Performance Guideline Verification . . . . .	100
<i>Jesper Larsson Träff</i>	
OMB-GPU: A Micro-Benchmark Suite for Evaluating MPI Libraries on GPU Clusters . . . . .	110
<i>D. Bureddy, H. Wang, A. Venkatesh, S. Pothuri, and D.K. Panda</i>	
Micro-applications for Communication Data Access Patterns and MPI Datatypes . . . . .	121
<i>Timo Schneider, Robert Gerstenberger, and Torsten Hoeﬂer</i>	

## Programming Models and New Architectures

Leveraging MPI's One-Sided Communication Interface for Shared-Memory Programming . . . . .	132
<i>Torsten Hoeﬂer, James Dinan, Darius Buntinas, Pavan Balaji, Brian W. Barrett, Ron Brightwell, William Gropp, Vivek Kale, and Rajeev Thakur</i>	
Wait-Free Message Passing Protocol for Non-coherent Shared Memory Architectures . . . . .	142
<i>Isaías A. Comprés Ureña, Michael Gerndt, and Carsten Trinitis</i>	

## Run-Time Support

An Efficient Kernel-Level Blocking MPI Implementation . . . . .	153
<i>Atsushi Hori, Toyohisa Kameyama, Yuichi Tsujita, Mitaro Namiki, and Yutaka Ishikawa</i>	
Automatic Resource-Centric Process Migration for MPI . . . . .	163
<i>Amnon Barak, Alexander Margolin, and Amnon Shiloh</i>	

An Integrated Runtime Scheduler for MPI .....	173
<i>Humaira Kamal and Alan Wagner</i>	

## Fault-Tolerance

High Performance Checksum Computation for Fault-Tolerant MPI over Infiniband.....	183
<i>Alexandre Denis, Francois Trahay, and Yutaka Ishikawa</i>	
An Evaluation of User-Level Failure Mitigation Support in MPI .....	193
<i>Wesley Bland, Aurelien Bouteiller, Thomas Herault, Joshua Hursey, George Bosilca, and Jack J. Dongarra</i>	

## Message-Passing Algorithms

Efficient MPI Implementation of a Parallel, Stable Merge Algorithm....	204
<i>Christian Siebert and Jesper Larsson Träff</i>	
Efficient Distributed Computation of Maximal Exact Matches .....	214
<i>Mohamed Abouelhoda and Sondos Seif</i>	
Scalable Algorithms for Constructing Balanced Spanning Trees on System-Ranked Process Groups .....	224
<i>Akhil Langer, Ramprasad Venkataraman, and Laxmikant Kale</i>	

## Message-Passing Applications

A Hybrid Parallelization of Air Quality Model with MPI and OpenMP.....	235
<i>Gian Franco Marras, Camillo Silibello, and Giuseppe Calori</i>	

## IMUDI: Improving MPI User and Developer Interaction

2 <sup>nd</sup> Special Session on Improving MPI User and Developer Interaction (IMUDI 2012) .....	246
<i>Dries Kimpe and Jason Cope</i>	
A Wish List for Efficient Adjoints of One-Sided MPI Communication...	248
<i>Michel Schanen and Uwe Naumann</i>	
On the Usability of the MPI Shared File Pointer Routines.....	258
<i>Mohamad Chaarawi, James Dinan, and Dries Kimpe</i>	
Extending MPI to Better Support Multi-application Interaction.....	268
<i>Jay Lofstead and Jai Dayal</i>	

Versatile Communication Algorithms for Data Analysis . . . . .	275
<i>Tom Peterka and Robert Ross</i>	

## Posters

High Performance Concurrent Multi-Path Communication for MPI . . . . .	285
<i>Rashid Hassani, Abbas Malekpour, Amirreza Fazely, and Peter Luksch</i>	
Improving Collectives by User Buffer Relocation . . . . .	287
<i>Juan Antonio Rico Gallego, Juan Carlos Díaz Martín, Carolina Gómez-Tostón Gutiérrez, and Álvaro Cortés Fácila</i>	
Asynchronous Checkpointing by Dedicated Checkpoint Threads . . . . .	289
<i>Faisal Shahzad, Markus Wittmann, Thomas Zeiser, and Gerhard Wellein</i>	
Verification of MPI Programs Using Session Types . . . . .	291
<i>Kohei Honda, Eduardo R.B. Marques, Francisco Martins, Nicholas Ng, Vasco T. Vasconcelos, and Nobuko Yoshida</i>	
Runtime Support for Adaptive Resource Provisioning in MPI Applications . . . . .	294
<i>Gonzalo Martín, David E. Singh, Maria-Cristina Marinescu, and Jesús Carretero</i>	
Revisiting Persistent Communication in MPI . . . . .	296
<i>Yutaka Ishikawa, Kengo Nakajima, and Atsushi Hori</i>	
StarPU-MPI: Task Programming over Clusters of Machines Enhanced with Accelerators . . . . .	298
<i>Cédric Augonnet, Olivier Aumage, Nathalie Furmento, Raymond Namyst, and Samuel Thibault</i>	
<b>Author Index . . . . .</b>	<b>301</b>