Data-Intensive Computing

The world is awash with digital data from social networks, blogs, business, science, and engineering. Data-intensive computing facilitates understanding of complex problems that must process massive amounts of data. Through the development of new classes of software, algorithms, and hardware, data-intensive applications can provide timely and meaningful analytical results in response to exponentially growing data complexity and associated analysis requirements. This emerging area brings many challenges that are different from traditional high-performance computing.

This reference for computing professionals and researchers describes the dimensions of the field, the key challenges, the state of the art, and the characteristics of likely approaches that future data-intensive problems will require. Chapters cover general principles and methods for designing such systems and for managing and analyzing the big data sets of today that live in the cloud, and describe example applications in bioinformatics and cyber-security that illustrate these principles in practice.

IAN GORTON is a Laboratory Fellow in Computational Sciences and Math at Pacific Northwest National Laboratory (PNNL), where he manages the Data Intensive Scientific Computing group and was the Chief Architect for PNNL's Data Intensive Computing Initiative. Gorton is a Senior Member of the IEEE Computer Society and a Fellow of the Australian Computer Society.

DEBORAH K. GRACIO joined Pacific Northwest National Laboratory in 1990 and is currently the Director of the Computational and Statistical Analytics Division and of the Data Intensive Computing Research Initiative. Since joining the laboratory, she has led the research, development, and management of multiple cross-disciplinary, multi-laboratory projects focused in the basic sciences and national security sectors.

Data-Intensive Computing

Architectures, Algorithms, and Applications

Edited by

IAN GORTON
Pacific Northwest National Laboratory

DEBORAH K. GRACIO
Pacific Northwest National Laboratory

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town,
Singapore, São Paulo, Delhi, Mexico City

Cambridge University Press
32 Avenue of the Americas, New York, NY 10013-2473, USA

www.cambridge.org
Information on this title: www.cambridge.org/9780521191951

First published 2013

Printed in the United States of America

A catalog record for this publication is available from the British Library.

Library of Congress Cataloging in Publication Data

Data-intensive computing : architectures, algorithms, and applications / [edited by]
Ian Gorton, Deborah K. Gracio.
pages cm
Includes bibliographical references and index.
ISBN 978-0-521-19195-1
1. High performance computing. 2. Database management. 3. Computer storage devices.
4. Software architecture. 5. Data transmission systems. I. Gorton, Ian.
II. Gracio, Deborah K., 1965–
QA76.88.D38 2012
004.5–dc23 2012015720

ISBN 978-0-521-19195-1 Hardback

Contents

List of Contributors

Ghaleb Abdulla Lawrence Livermore National Laboratory

Justin Almquist Pacific Northwest National Laboratory

Magdalena Balazinska University of Washington

Jacek Becla Stanford University

Daniel M. Best Pacific Northwest National Laboratory

Shawn J. Bohn Pacific Northwest National Laboratory

Daniel Chavarría-Miranda Pacific Northwest National Laboratory

Terence Critchlow Pacific Northwest National Laboratory

Scott Dowson Pacific Northwest National Laboratory

Ian Gorton Pacific Northwest National Laboratory

Deborah K. Gracio Pacific Northwest National Laboratory

Wes Hatley Future Point Systems

Bill Howe University of Washington

Chandrika Kamath Lawrence Livermore National Laboratory

Sam Lang Pacific Northwest National Laboratory

Anna Liu National ICT Australia (NICTA), University of New South Wales

Kerstin Kleese-Van Dam Pacific Northwest National Laboratory

Douglas V. Love Pacific Northwest National Laboratory

Lee Ann McCue Pacific Northwest National Laboratory

Jason McDermott Pacific Northwest National Laboratory

Deborah L. McGuinness Rensselaer Polytechnic Institute

Patrick Nichols Pacific Northwest National Laboratory

Christopher Oehmen Pacific Northwest National Laboratory

William A. Pike Pacific Northwest National Laboratory

Sherif Sakr National ICT Australia (NICTA), University of New South Wales

Antonino Tumeo Pacific Northwest National Laboratory

Oreste Villa Pacific Northwest National Laboratory

Bobbie-Jo Webb-Robertson Pacific Northwest National Laboratory