INPUT/OUTPUT IN PARALLEL AND DISTRIBUTED COMPUTER SYSTEMS

THE KLUWER INTERNATIONAL SERIES
IN ENGINEERING AND COMPUTER SCIENCE

INPUT/OUTPUT IN PARALLEL AND DISTRIBUTED COMPUTER SYSTEMS

edited by

Ravi Jain
Bell Communications Research
Morristown,New Jersey, USA

John Werth
University of Texas at Austin
Austin, Texas, USA

James C. Browne
University of Teas at Austin
Austin, Texas, USA

KLUWER ACADEMIC PUBLISHERS
Boston / Dordrecht / London

Distributors for North America:
Kluwer Academic Publishers
101 Philip Drive
Assinippi Park
Norwell, Massachusetts 02061 USA

Distributors for all other countries:
Kluwer Academic Publishers Group
Distribution Centre
Post Office Box 322
3300 AH Dordrecht, THE NETHERLANDS

Library of Congress Cataloging-in-Publication Data

A C.I.P. Catalogue record for this book is available
from the Library of Congress.

ISBN-13: 978-1-4612-8607-3 e-ISBN-13: 978-1-4613-1401-1
DOI: 10.1007/978-1-4613-1401-1

To

Meera, Laurie and Gayle

CONTENTS

PREFACE

I/O for parallel and distributed computer systems has drawn increasing attention over the last few years as it has become apparent that I/O performance, rather than CPU performance, may be the key limiting factor in the performance of future systems. This *I/O bottleneck* is caused by the increasing speed mismatch between processing units and storage devices, the use of multiple processors operating simultaneously in parallel and distributed systems, and by the increasing I/O demands of new classes of applications, like multimedia. It is also important to note that, to varying degrees, the I/O bottleneck exists at multiple levels of the memory hierarchy. All indications are that the I/O bottleneck will be with us for some time to come, and is likely to increase in importance.

These realizations prompted us to advocate that the I/O bottleneck be addressed systematically at all levels of parallel and distributed system design. Thus while there are solutions which focus on one aspect of the system (e.g., architectural solutions like RAID), we felt that the benefits of these solutions would not be realized unless I/O-efficient design was integrated into applications, algorithms, compilers, operating systems, and architectures for parallel and distributed systems. With this view we initiated, in 1993, a workshop dedicated to I/O in parallel and distributed systems, held in conjunction with the International Parallel Processing Symposium (IPPS). The workshop drew substantial interest, and has now become an annual event. Papers submitted to the workshop were refereed, and in 1994 and 1995 about 30%-40% of the submitted papers were accepted for presentation. In 1996 the workshop on I/O in Parallel and Distributed Systems (IOPADS) is being held as an independent workshop with the Federated Computing Research Conference, and continues to draw international research participation.

This book is divided into three parts. Part I, the Introduction, contains four invited chapters which provide a tutorial and survey of I/O issues in parallel and distributed systems. The chapters in Parts II and III contain selected research papers from the 1994 and 1995 IOPADS workshops; many of these papers have been substantially revised and updated for inclusion in this volume. Part II collects the papers from both years which dealt with various aspects of system software, and Part III those addressing primarily architectural issues.

The first chapter in Part I provides an overview of I/O issues, and surveys upcoming trends in this area, such as the convergence of networking and I/O, the increasing importance of the Internet and World-Wide Web as a new level of the system memory hierarchy, and the challenges posed by mobile and wireless computing. Chapter 2 by Shriver and Nodine presents an introduction to parallel I/O models and algorithms. Chapter 3 by Thakur and Choudhary, and Chapter 4 by Kotz, provide introductions to compiler issues and architectural approaches, respectively. We hope Part I will prove a useful source for graduate students and researchers new to the area.

The papers in Part II include those on compiler support, programming models, interfaces, file systems and scheduling. Chapter 5, by a team from IBM and NASA, presents the MPI-IO parallel I/O interface, and Chapter 6 by Thakur and Choudhary describes work on runtime support for out-of-core data parallel algorithms. Chapter 7 by Baylor et al presents a summary of the Vesta parallel file system developed at IBM, and its use for collecting parallel I/O workload characteristics, while Chapter 8 by Kalns and Hsu describes its use for an interesting and important application, video-on-demand. In Chapter 9, Nieuwejaar and Kotz present the results of tracing a parallel file system for scientific applications; the results indicate extensions needed to the interface provided to the programmer. Moyer and Sunderam, in Chapter 10, address the concurrency control issues that can arise when even a single read or write operation results in parallel I/O operations on multiple storage devices. In Chapter 11, Durand et al present distributed algorithms for scheduling parallel I/O operations so as to minimize their completion time. In Chapter 12, Sinclair et al describe scenarios in which multiple processes engaging in parallel I/O can conflict such that a subset of them monopolize the I/O resources; this work has implications for data allocation and task partitioning in parallel systems.

Part III of the book focuses on issues relating to system architecture. Chapter 13, by Baylor et al of IBM, describes a performance evaluation of the massively parallel I/O subsystem of the Vulcan MPP, with results on the placement of I/O nodes in the system. VanderLeest and Iyer present, in Chapter 14, a methodology for measuring bus contention, a critical resource in parallel I/O, and the results of such a study on a specific system. In Chapter 15, Schloss and Vernick present a Hybrid Client-Server Architecture, in which the traditional client-server architecture is modified to give clients access to both the server and its disks. In Chapter 16, Yokota and Mimatsu propose Data-reconstruction networks for I/O subsystems, where each node has a set of disks and nodes are interconnected by a network separate from the primary processor interconnection network. Finally, in Chapter 17, Asthana et al describe an I/O subsystem in which processing logic is associated with each memory chip, offloading some low-level I/O-related tasks from the CPU and thus speeding up overall system operation.

The depth and breadth of the chapters in Parts II and III indicates the vitality of this fast-growing research area, and we hope they will stimulate further study and integration of approaches for alleviating the I/O bottleneck.

We would like to thank the members of the IOPADS 1994 and 1995 program committees who reviewed papers. The program committee members were: Abhaya Asthana (AT&T Bell Labs), Larry Berdahl (Lawrence Livermore), Peter Chen (Univ. of Michigan), Alok Choudhary (Syracuse), Peter Corbett (IBM Watson), Tom Cormen (Dartmouth), David DeWitt (Univ. of Wisconsin), Sam Fineberg (NASA Ames), Shahram Ghandeharizadeh (USC), Paul Messina (Caltech), John Nickolls (MasPar), and Wayne Roiger (Cray Research). We also thank the many other reviewers who provided us with reviews of the submitted papers. Papers from the 1994 Workshop also appeared, in condensed form, in the ACM SIGARCH newsletter "Computer Architecture News", Oct. 1994; thanks are due to Doug DeGroot, editor, for his help in that regard. We thank Scott Delman of Kluwer for his help and patience during the long process of preparing this book, and Ravi Jain would like to thank Michael Kramer of Bellcore for his support.

Finally, we would like to thank the organizers of IPPS, and especially IPPS Chair Viktor Prasanna, for the opportunity to hold the IOPADS workshops in order to bring together researchers addressing the I/O issues in parallel and distributed systems.

Ravi Jain, John Werth and J. C. Browne